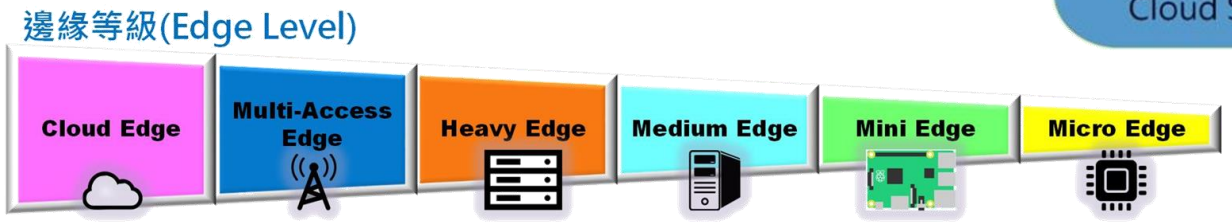
國立臺灣科技大學
NATIONAL TAIWAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Department of Computer Science & Information Engineering
資 訊 工 程 系

人工智慧與邊緣運算實務

02 邊緣運算硬體



資訊工程系 許哲豪 助理教授

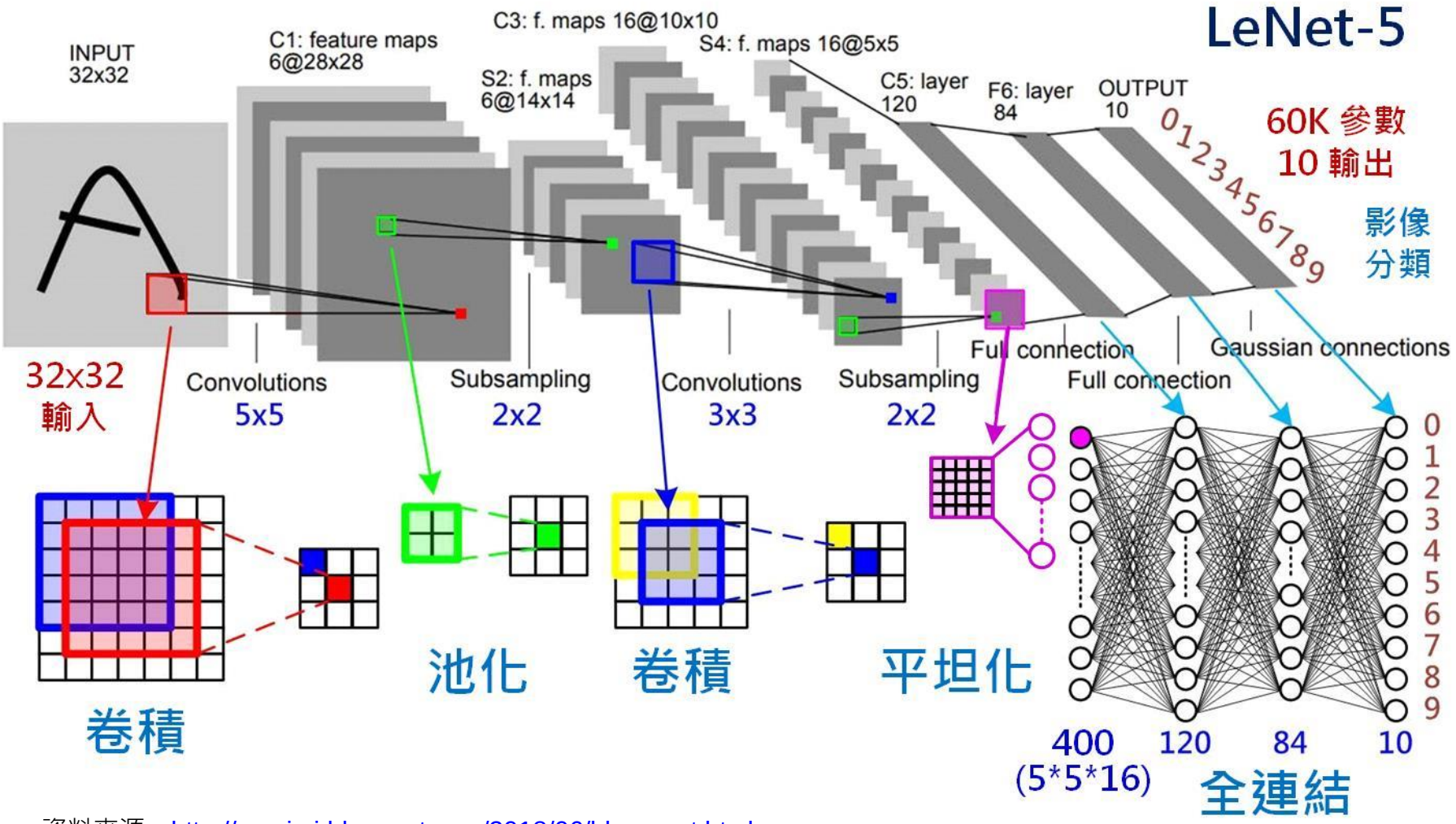
簡報大綱

- **2.1 基本運算原理**
  - 卷積神經網路
  - 數字表示系統
  - 矩陣 / 張量運算
  - 平行 / 並行運算
- **2.2 加速運算晶片**
  - CPU
  - GPU
  - FPGA
  - NPU
  - ASIC
  - 其它類型
- **2.3 開發板類型**
  - 單板微電腦
  - USB加速棒
  - 模組板卡
  - 整合型裝置
- **2.4 硬體選用評估**
  - 運算效能
  - 開發工具
  - 應用情境
  - 週邊擴展

2.1 基本運算原理

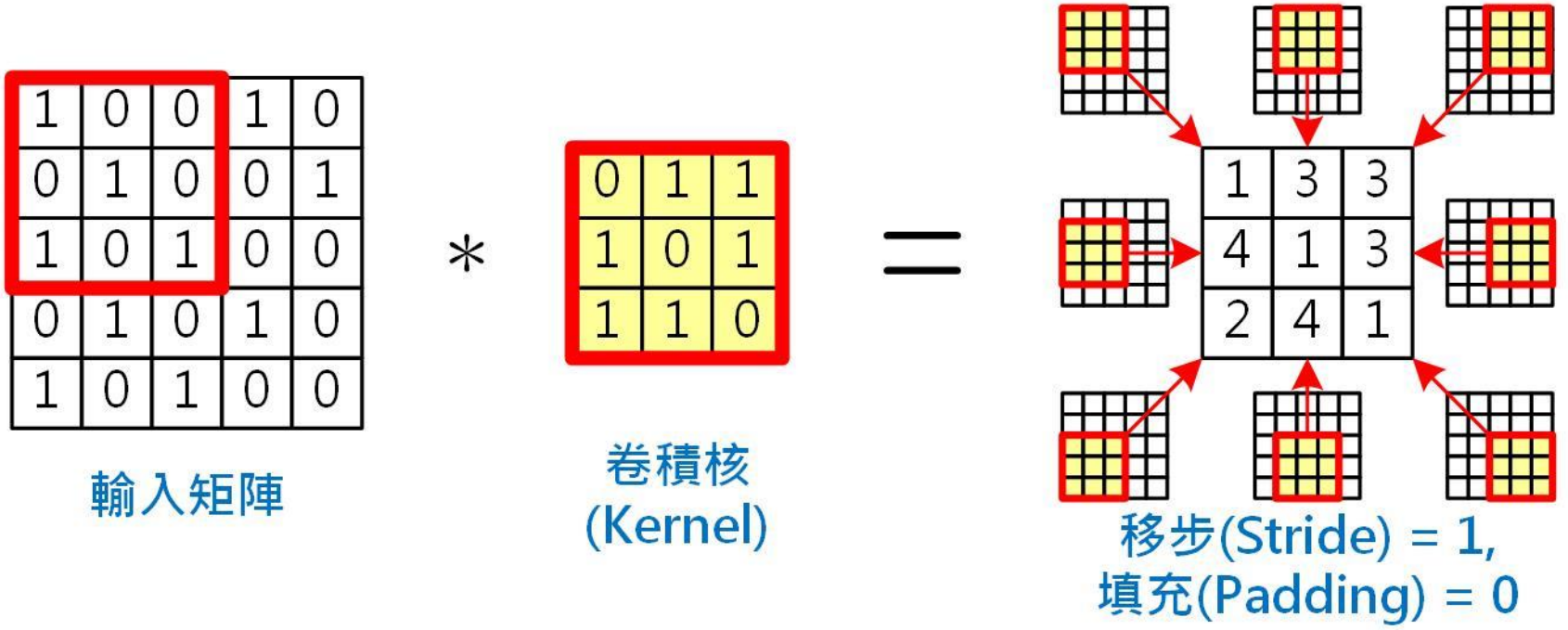
- 卷積神經網路
- 數字表示系統
- 矩陣 / 張量運算
- 平行 / 並行運算

2.1.1 卷積神經網路



資料來源：http://omnixri.blogspot.com/2018/06/blog-post.html

2.1.1 卷積神經網路—卷積



1	2	3
4	5	6
7	8	9

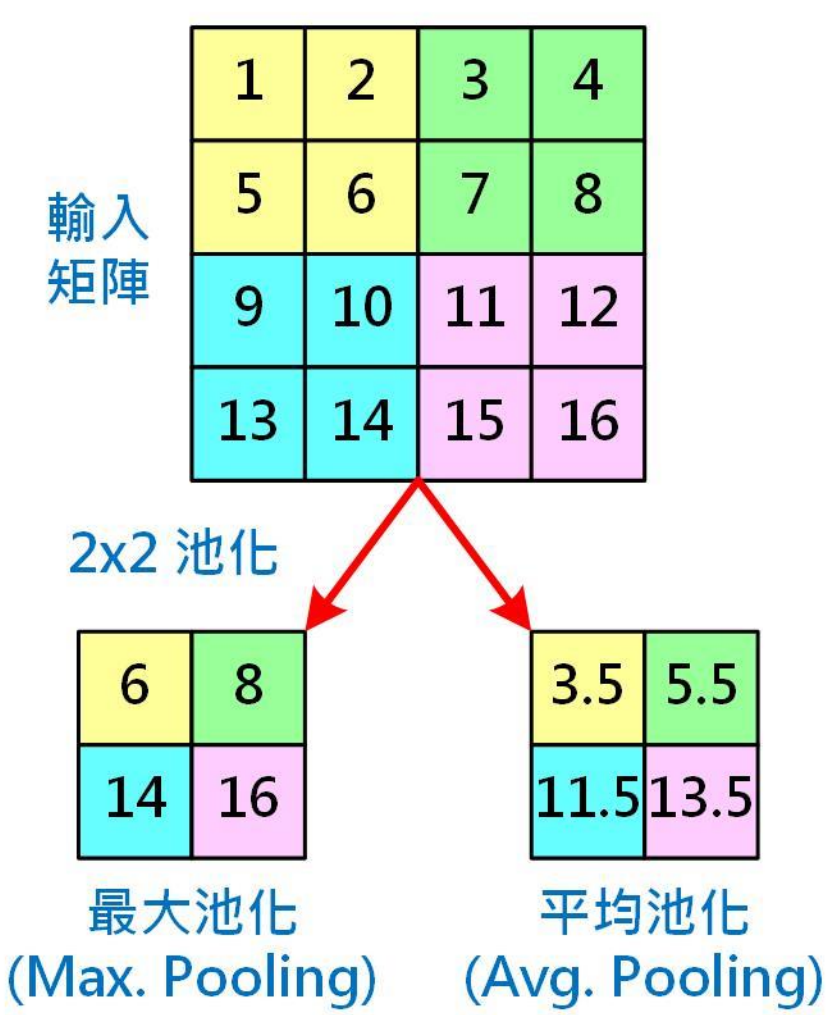
*

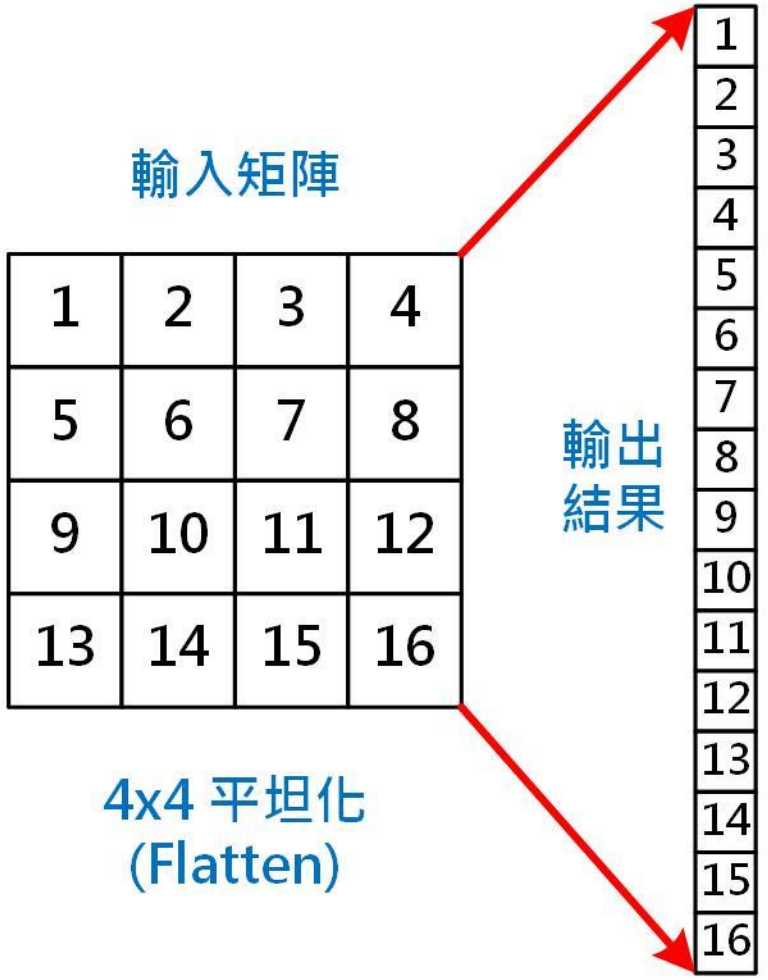
A	B	C
D	E	F
G	H	I

=

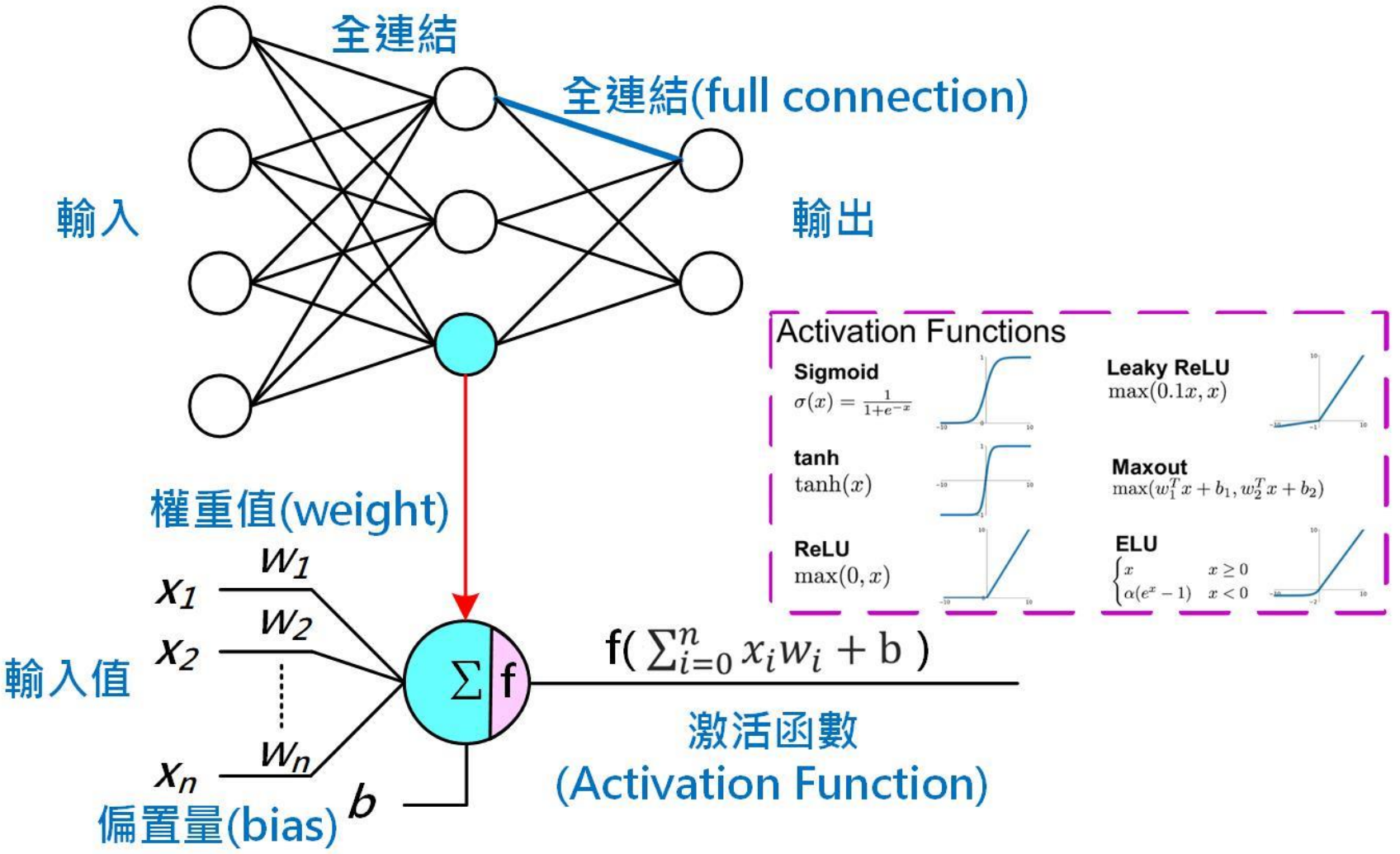
1*A + 2*B + 3*C +
4*D + 5*E + 6*F +
7*G + 8*H + 9*I

2.1.1 卷積神經網路—池化、平坦化

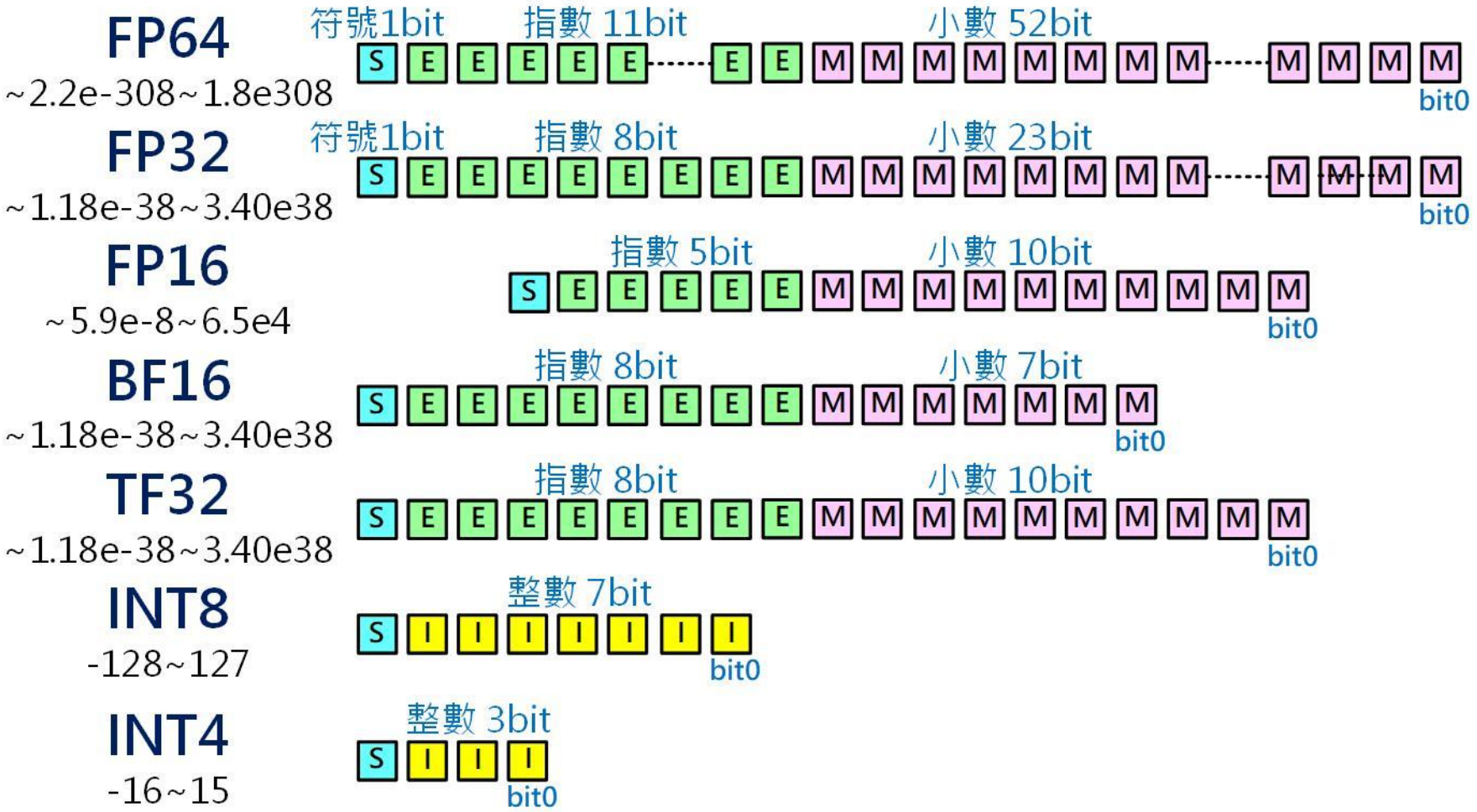




2.1.1 卷積神經網路—全連結

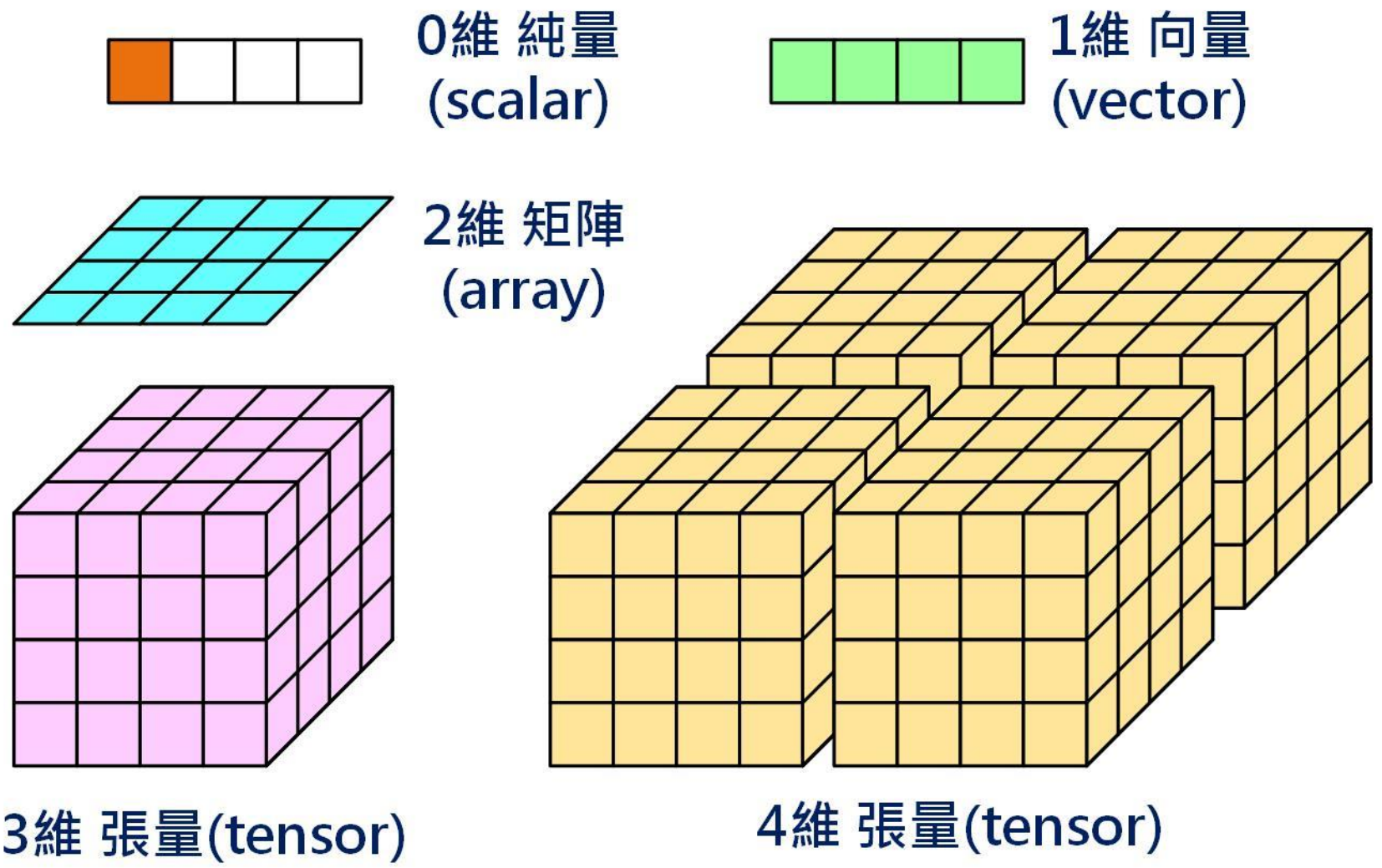


2.1.2 數字表示系統

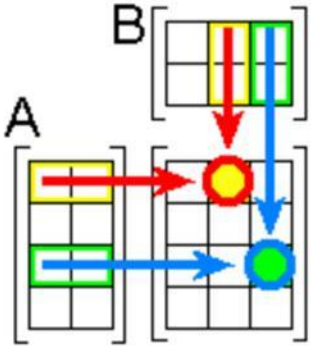


資料來源：https://moocaholic.medium.com/fp64-fp32-fp16-bfloat16-tf32-and-other-members-of-the-zoo-a1ca7897d407

2.1.3 矩陣 / 張量運算–名詞定義



2.1.3 矩陣 / 張量運算–矩陣乘法



$$(AB)_{ij} = \sum_{r=1}^{n} a_{ir} b_{rj} = a_{i1} b_{1j} + a_{i2} b_{2j} + \cdots + a_{in} b_{nj}$$

$$(AB)_{1,2} = \sum_{r=1}^{2} a_{1,r} b_{r,2} = a_{1,1} b_{1,2} + a_{1,2} b_{2,2}$$

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \cdots \\ a_{2,1} & a_{2,2} & a_{2,3} & \cdots \\ a_{3,1} & a_{3,2} & a_{3,3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} = \begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ \vdots \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} & \cdots \\ b_{2,1} & b_{2,2} & b_{2,3} & \cdots \\ b_{3,1} & b_{3,2} & b_{3,3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} = \begin{bmatrix} B_1 & B_2 & B_3 & \ldots \end{bmatrix}$$

$$\mathbf{AB} = \begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ \vdots \end{bmatrix} \times \begin{bmatrix} B_1 & B_2 & B_3 & \ldots \end{bmatrix} = \begin{bmatrix} (A_1 \cdot B_1) & (A_1 \cdot B_2) & (A_1 \cdot B_3) & \cdots \\ (A_2 \cdot B_1) & (A_2 \cdot B_2) & (A_2 \cdot B_3) & \cdots \\ (A_3 \cdot B_1) & (A_3 \cdot B_2) & (A_3 \cdot B_3) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

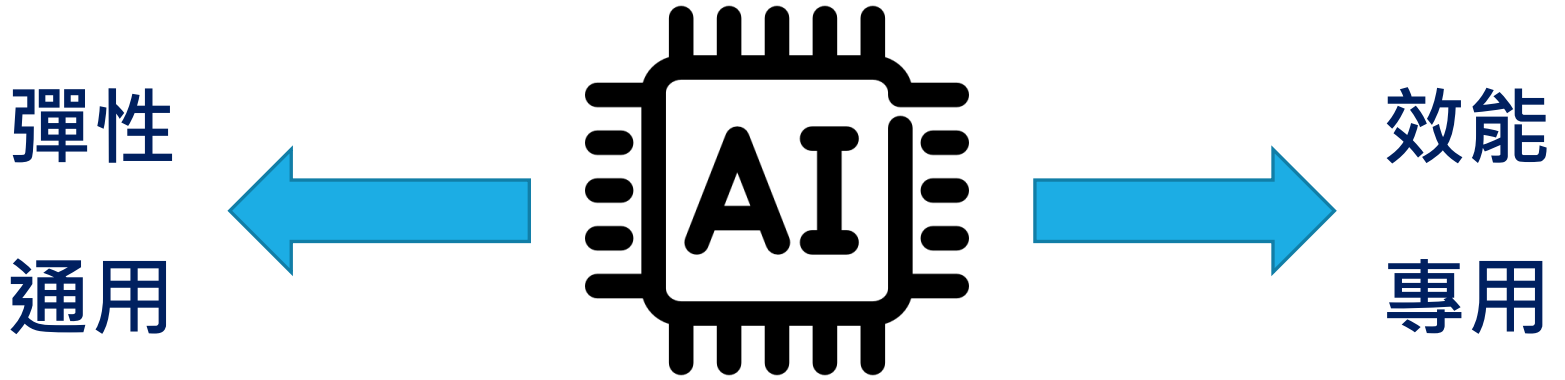
2.1.4 平行 / 並行運算

- **平行(Parallelism)**
  - 將工作分割成多份不同參數但內容相同工作內容，同時執行，待全部完成，再進行下一動作。
  - 如矩陣乘法是由很多個AxB+C組合而成，傳統上須依序計算再組合結果，此時就很適合分割成很多小工作一起計算，再合併結果得到所需結果。
- **並行(Concurrency)**
  - 可將工作(task)拆分成多個獨立工作，內容可不相同，可同時運行，互不干擾，亦可不同時完成。
  - 如編譯一個由很多函式組成的程式，利用多處理器（或多執行緒）方式同時編譯，待全部編譯完成後再連結成執行檔。

2.2 加速運算晶片

- CPU
- GPU
- FPGA
- NPU
- ASIC
- 其它類型

2.2 加速運算晶片—發展方向



多核心

高平行度

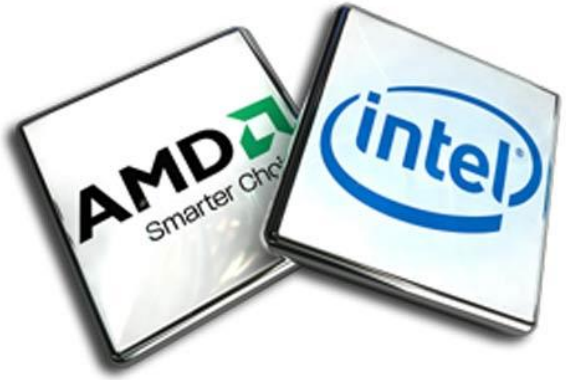
專用IC前驗證

神經網路專用

特定AI專用IC

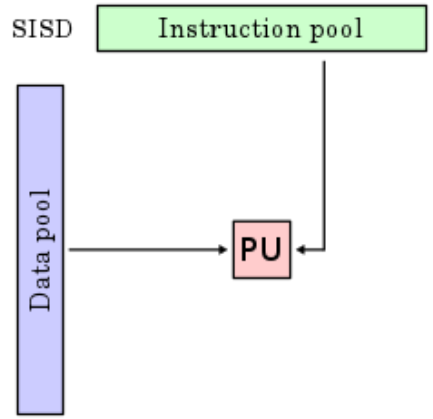
2.2.1 CPU—指令集架構(ISA)

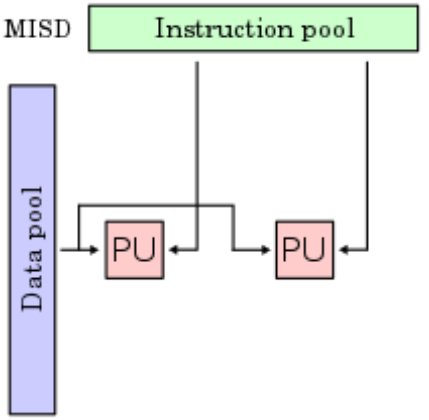
複雜指令集 (CISC)	精簡指令集 (RISC)
• INTEL(x86) • AMD • VIA	• ARM • RISC-V • MIPS

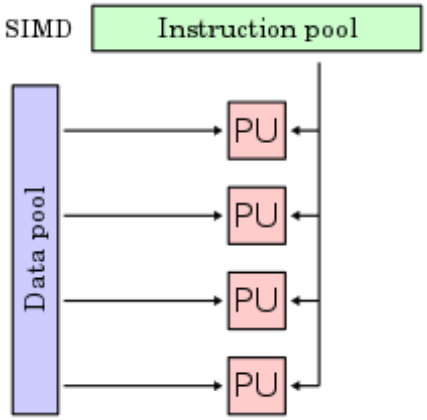


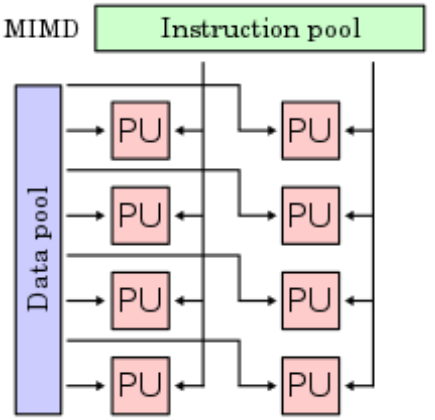
arm RISC-V MIPS

2.2.1 CPU—指令流與資料流







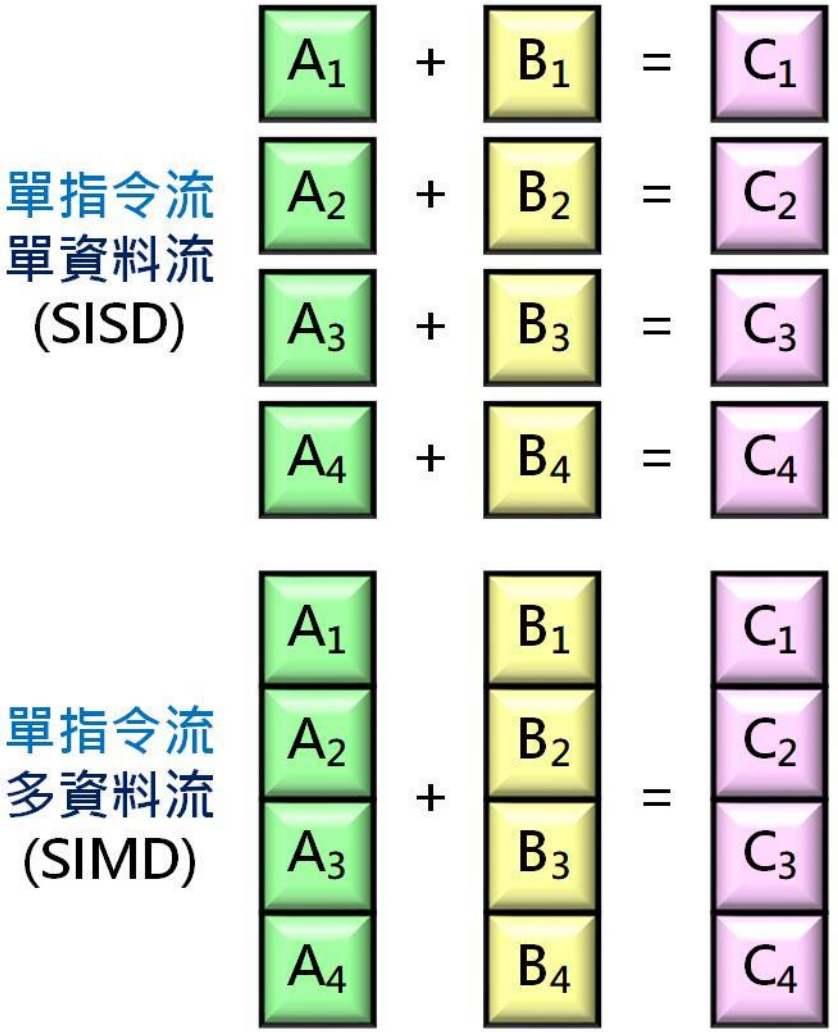


2.2.1 CPU–x86 SIMD

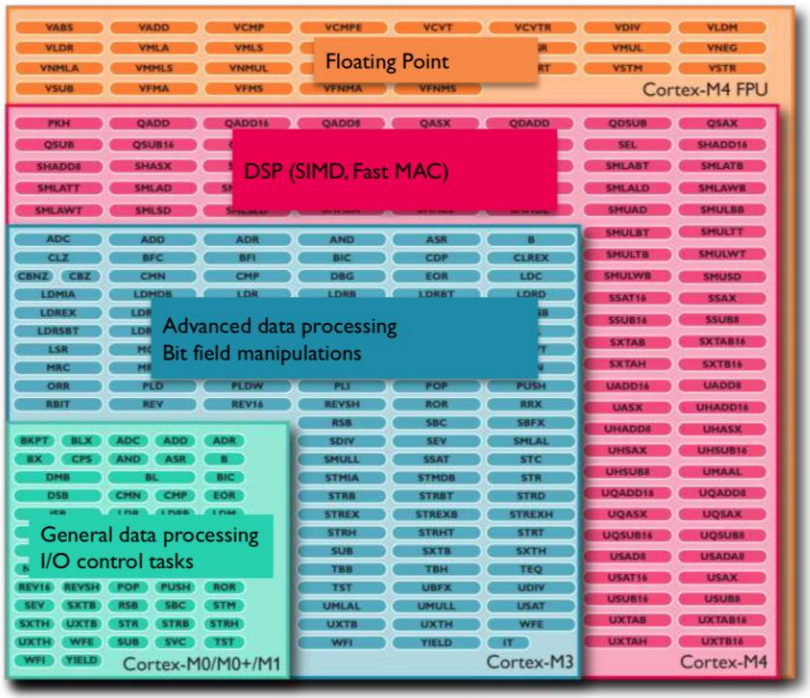
MMX
- 8x64bit暫存器
- 暫存器可作為整數 8x8bit,4x16bit, 2x32bit

SSE
- 加入浮點數
- 128bit指令長度
- SSE2 ~ SSE5

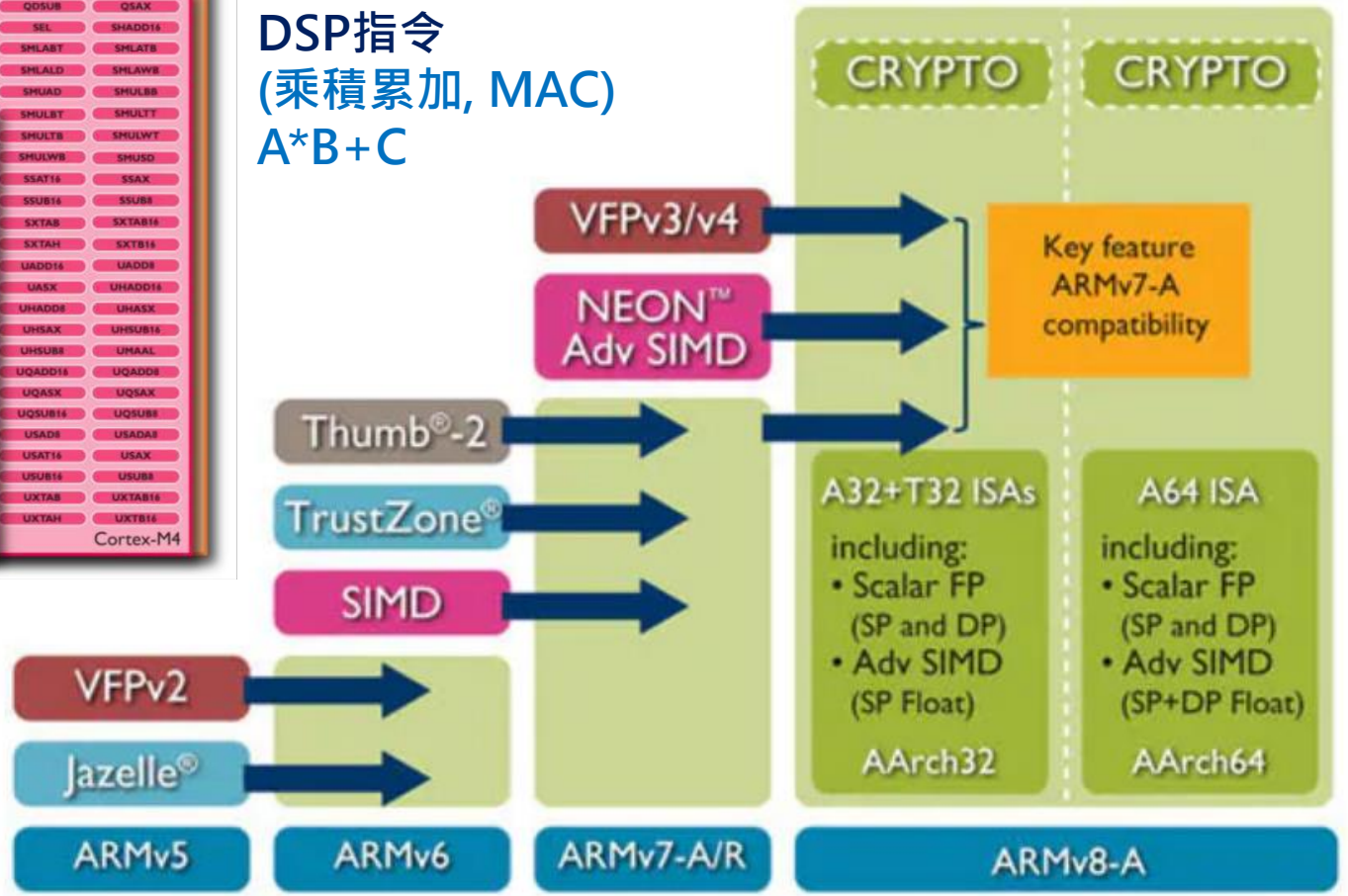
AVX
- AVX/ AVX2 (256bit)
- 512bit指令長度



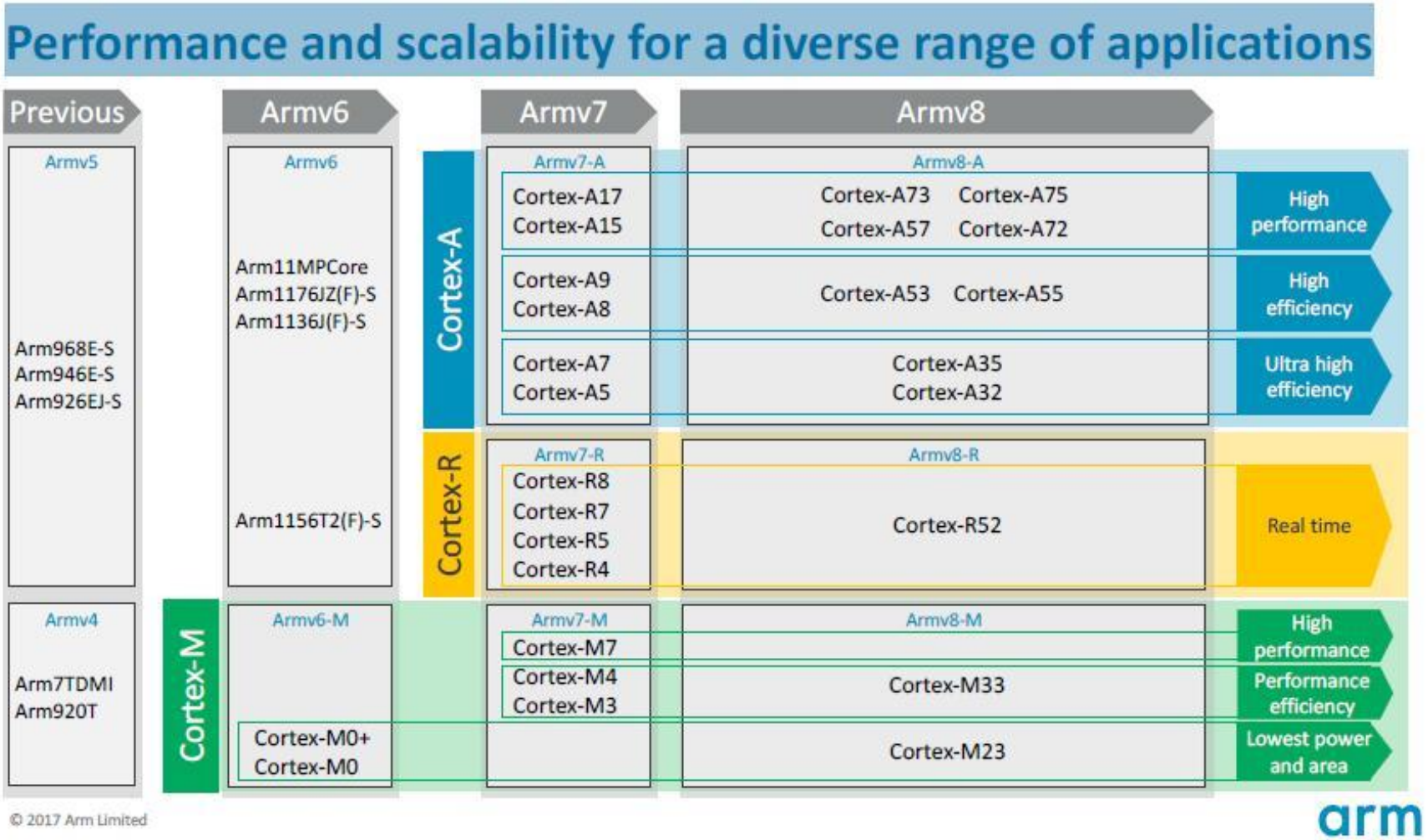
2.2.1 CPU–ARM NEON



DSP指令
(乘積累加, MAC)
A*B+C

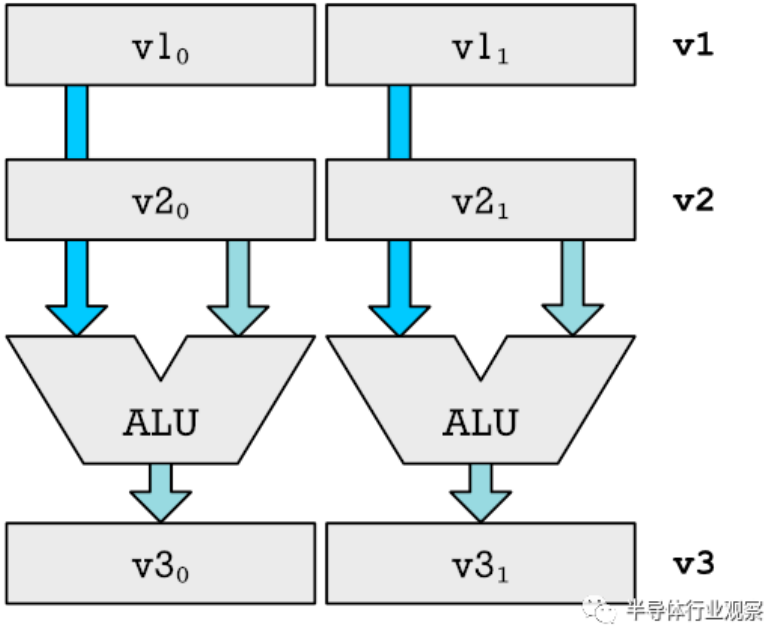


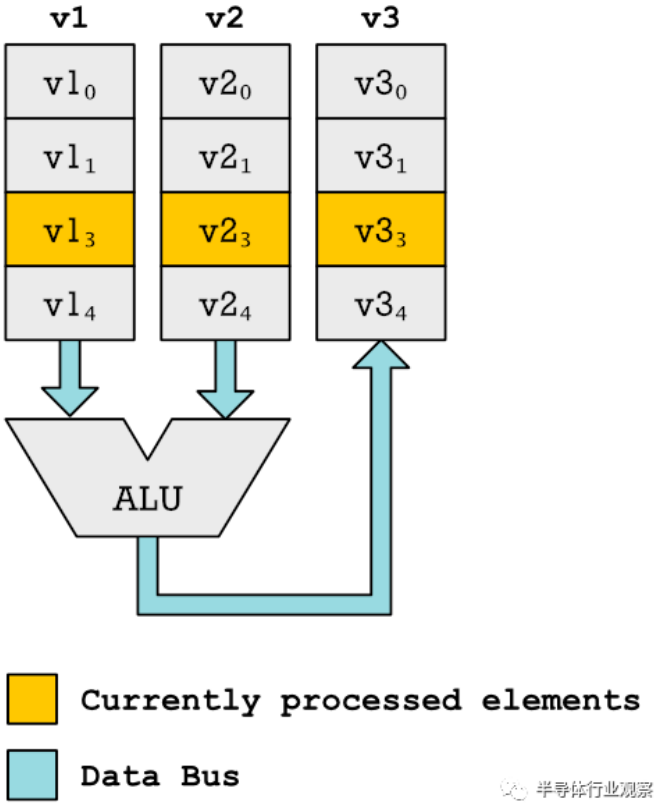
2.2.1 CPU–ARM家族



2.2.1 CPU—RISC-V向量指令集

- 擴充指令集 V ：向量運算
- 指充指令集 P ：單指令多資料流(SIMD)

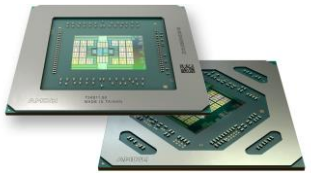




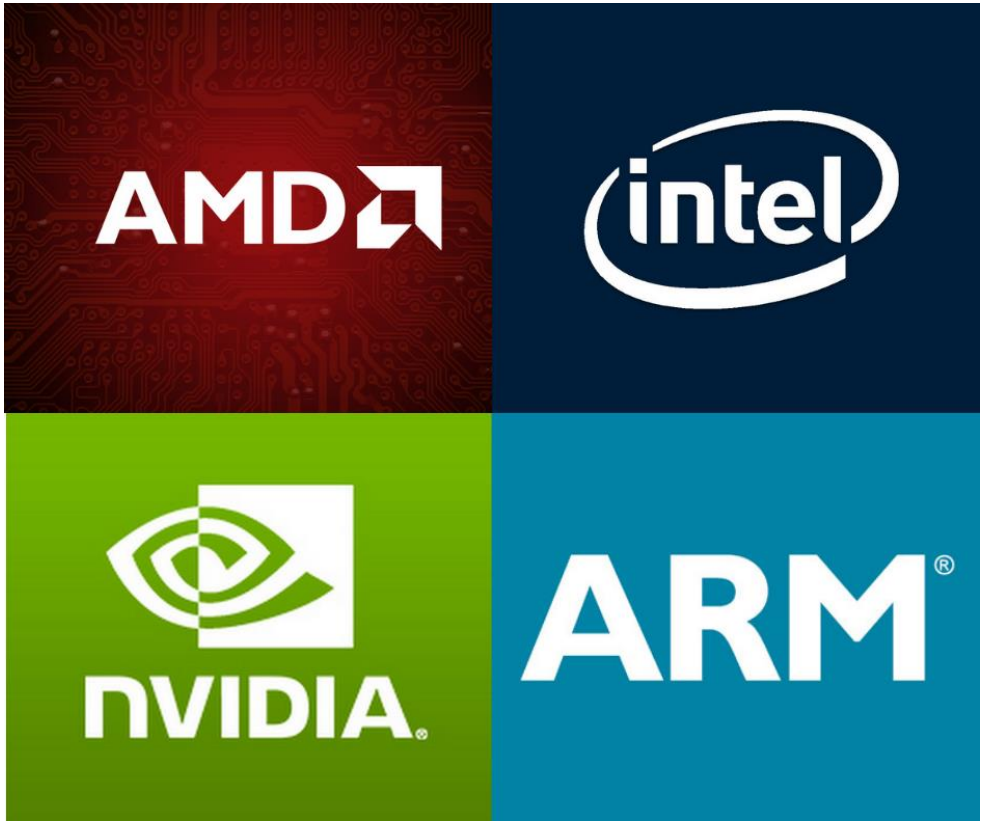
資料來源：https://www.riscv-mcu.com/article-show-id-541.html

2.2.2 GPU—主要供應商

Radeon



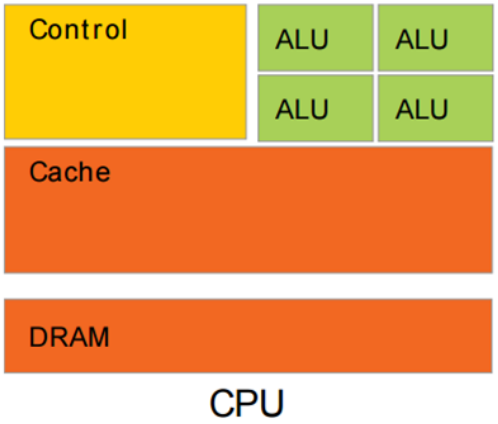
GeForce
Quadro/Tesla
Jetson/DGX

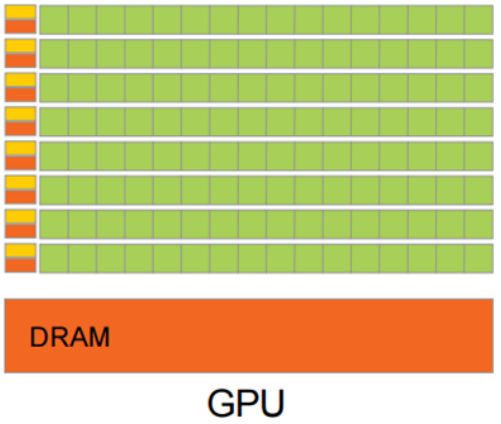


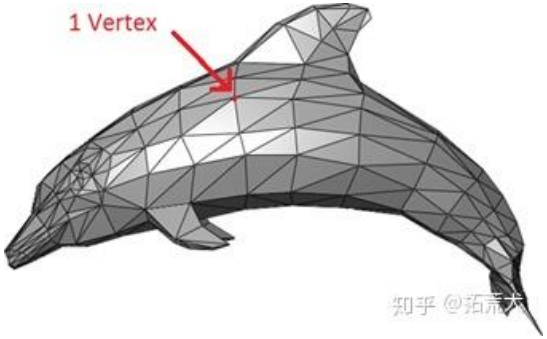
HD Graphic
Iris Xe

MALI

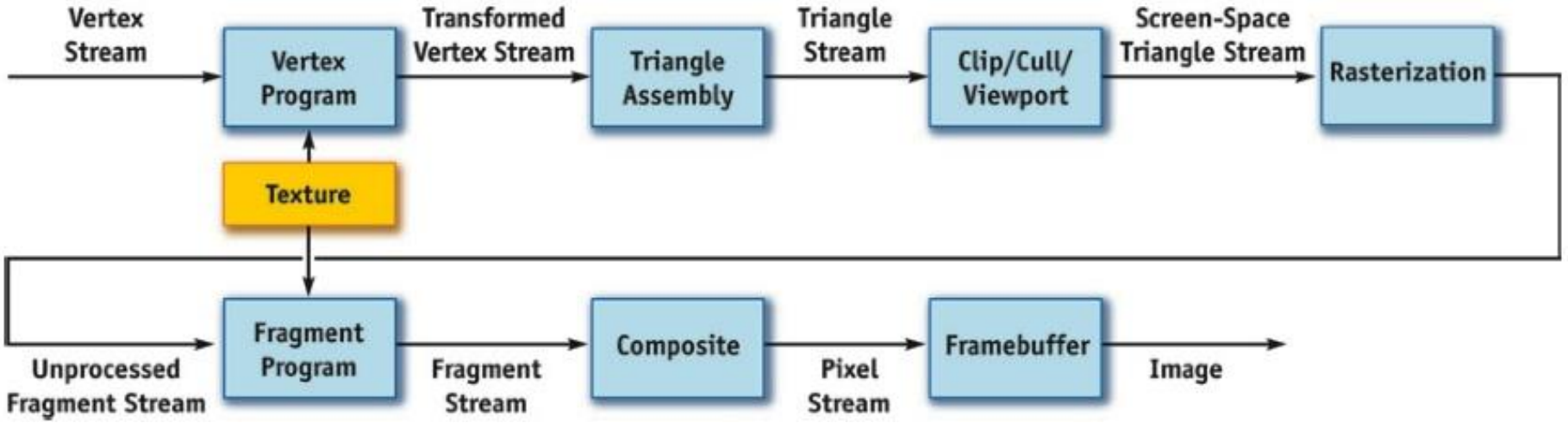
2.2.2 GPU—結構與工作流程



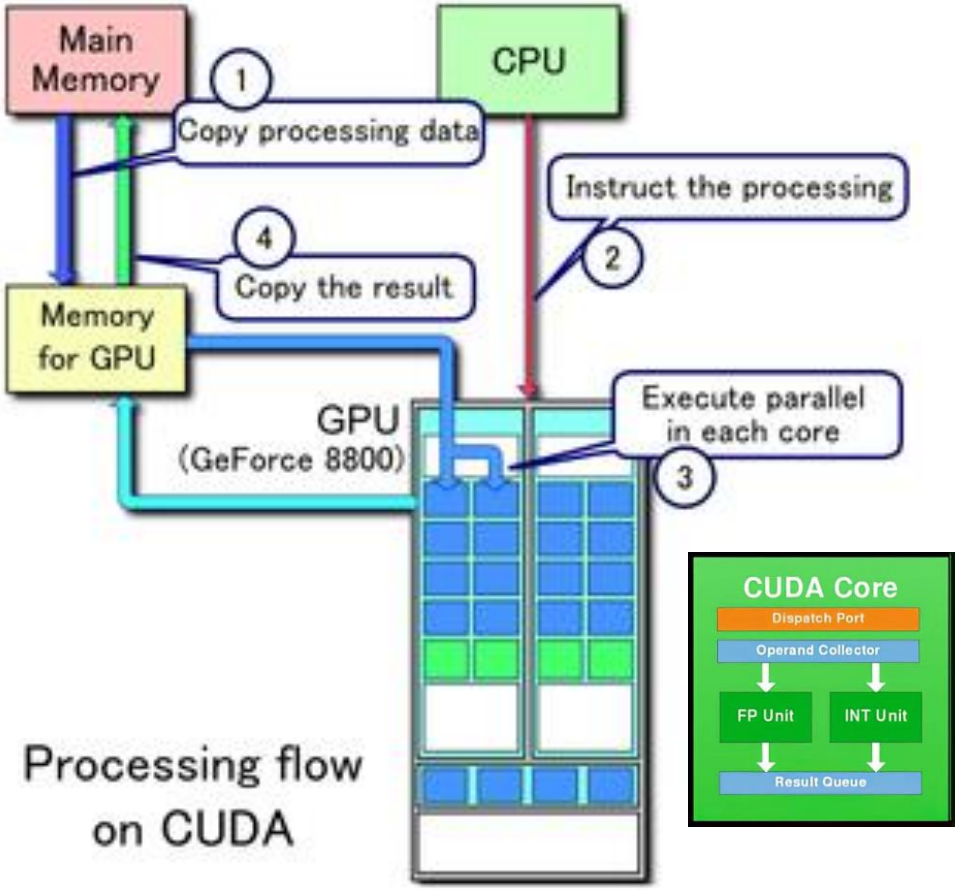




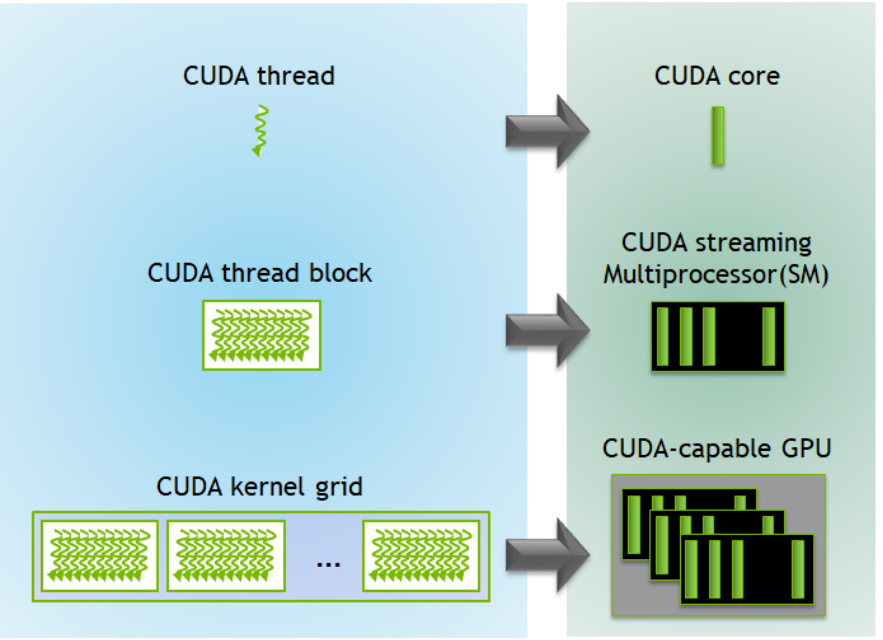
渲染(Render)



2.2.2 GPU–Nvidia Cuda Core

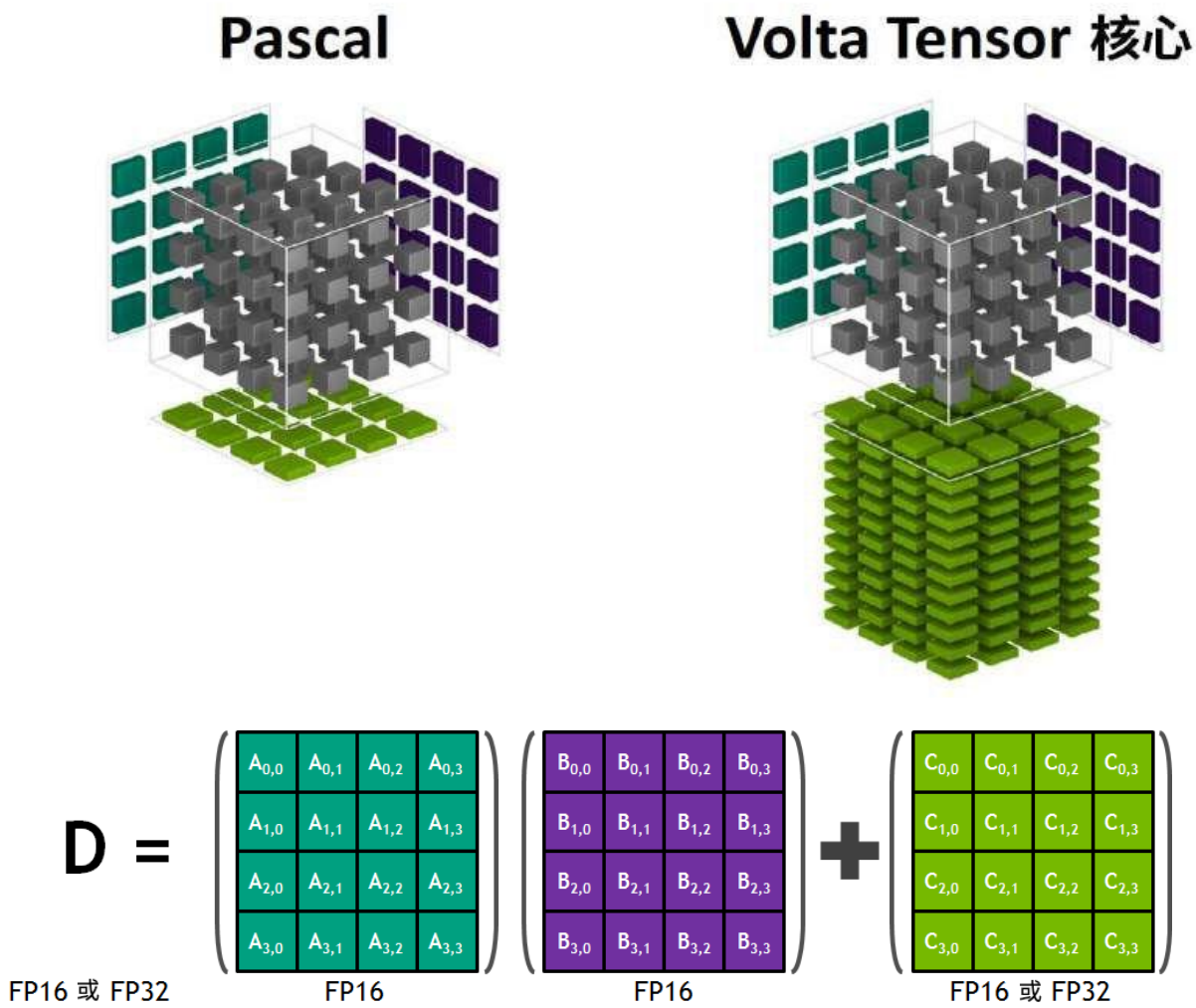


- 計算架構
  - GPGPU / CUDA / OpenCL



資料來源：https://zh.wikipedia.org/wiki/CUDA

2.2.2 GPU–Nvidia Tensor Core



2.2.2 GPU—Cuda / Tensor Core

訓練用

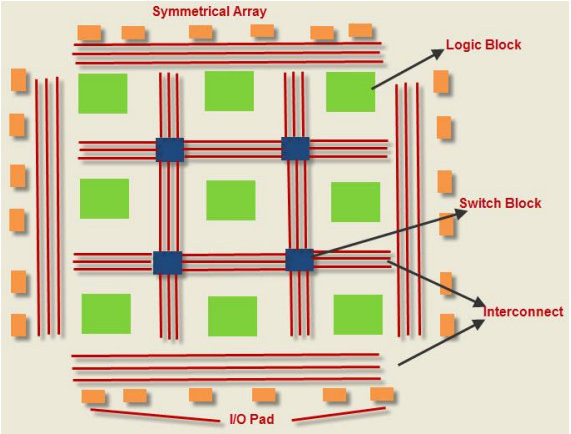
NVIDIA GPU	A100	TU102	GV100	GP100
GPU架構	Ampere	Turing	Volta	Pascal
製程	TSMC 7nm	TSMC 12nm	TSMC 12nm	TSMC 16nm
晶片面積	826mm2	754mm2	815mm2	610mm2
電晶體數目	542億	186億	211億	153億
FP32 CUDA Cores	6912	4608	5120	3584
GPU Boost Clock	1.41GHz	1770MHz	1530MHz	1480MHz
記憶體時脈	2.4Gbps 5120bit	14Gbps 384bit	1.75Gbps 4096bit	1.4Gbps 4096bit
記憶體	40GB HBM2	24GB GDDR6	16/32GB HBM2	16GB HBM2
Single Precision (FP32)	19.5 TFLOPs	16.3 TFLOP	15.7 TFLOPs	10.6 TFLOPs
INT8 Tensor	624 TOPs	261 TOPs	N/A	N/A
FP16 Tensor	312 TFLOPs	130.5 TFLOPs	125 TFLOPs	N/A
TF32 Tensor	156 TFLOPs	N/A	N/A	N/A
TDP	400W	260W	300/350W	300W

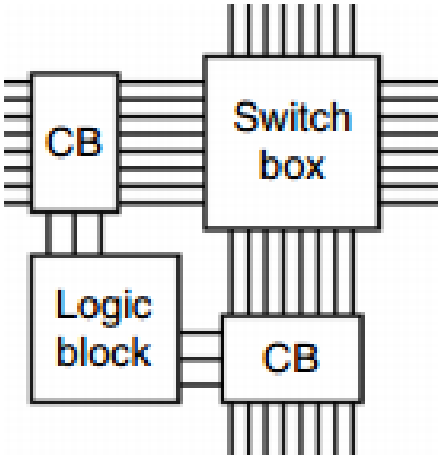
推論用

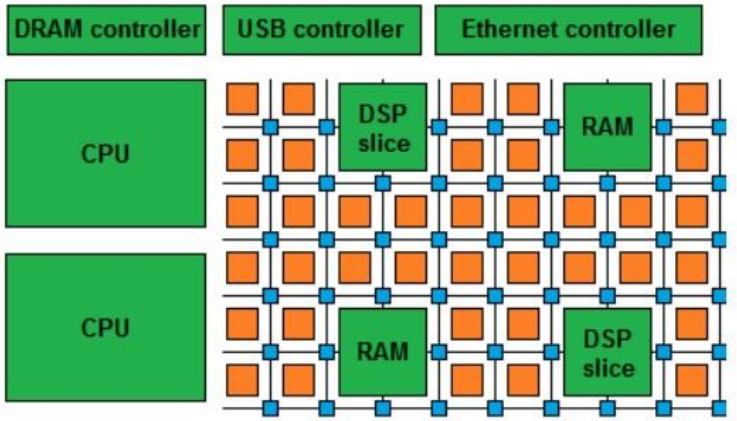
Jetson Nano	Jetson TX2	Jetson Xavier NX	Jetson AGX Xavier
NVIDIA Maxwell 架構，具有128個 NVIDIA CUDA核心	NVIDIA Pascal 架構，具有256個 NVIDIA CUDA核心	NVIDIA Volta架構 具有384個 NVIDIA CUDA核心 和48個Tensor核心	NVIDIA Volta架構 具有512個 NVIDIA CUDA核心 和64個Tensor核心

2.2.3 FPGA

現場可程式化邏輯陣列(Field - Programmable Gate Array)







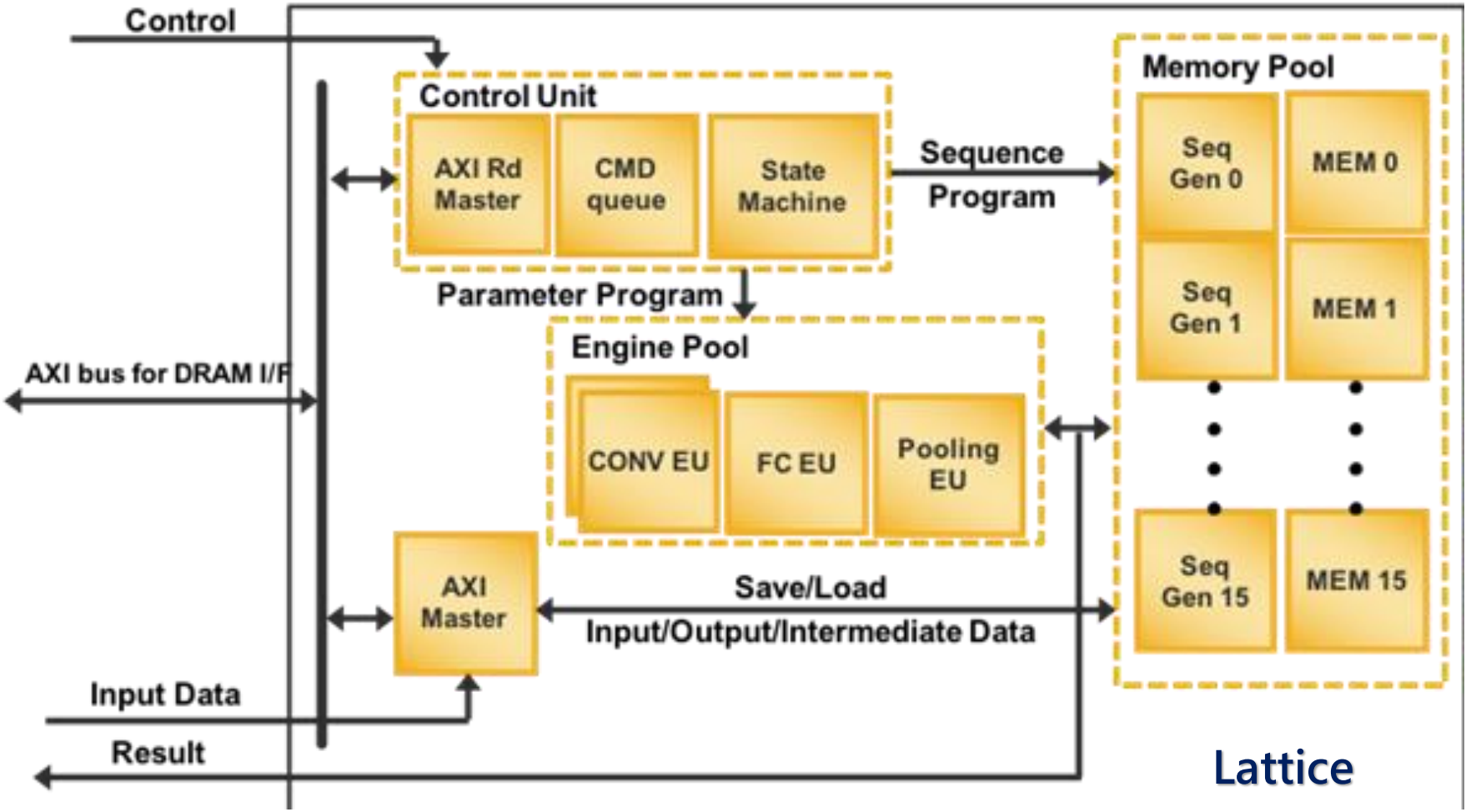
Xilinx

Intel(Altera)

Microsemi

Lattice

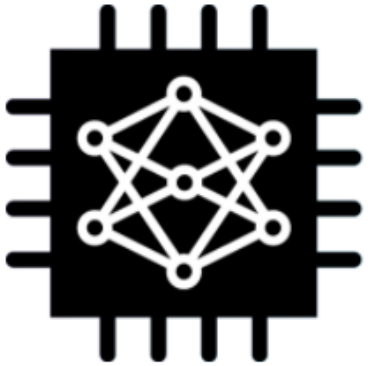
2.2.3 FPGA–CNN加速器



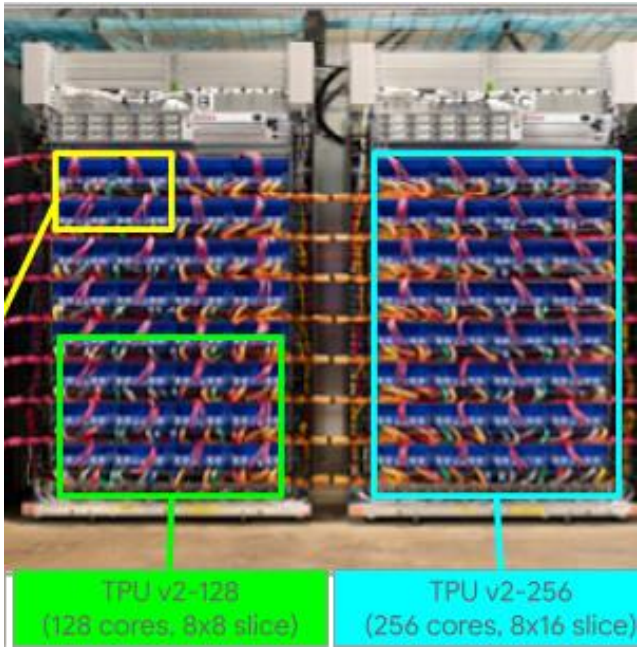
資料來源：https://www.latticesemi.com/products/designsoftwareandip/intellectualproperty/ipcore/ipcores04/cnn

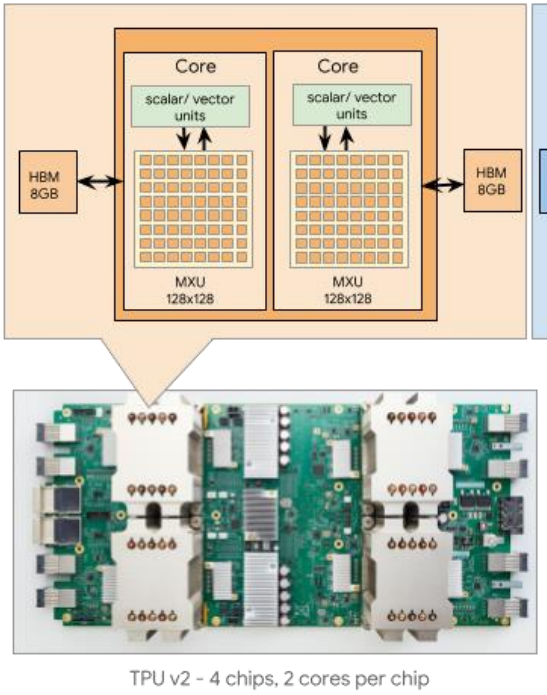
2.2.4 NPU

- 以神經網路運算為主要工作的晶片，屬於ASIC的一種，各廠家稱呼方式略有不同。
- 可支援通用型開發框架及模型，如ONNX, TensorFlow, PyTorch, Caffe等。
- 常見相關產品
  - Google TPU
  - Intel(Movidius) MA2485
  - 耐能Kneron KL520
  - ARM Ethos U55
  - 嘉楠勘智 K210
  - 瑞芯微 RK1808

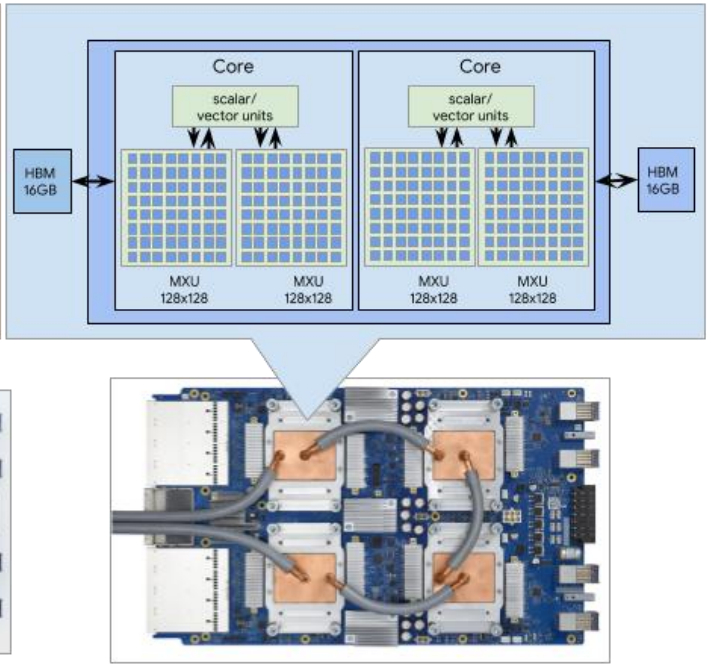


2.2.4 NPU—Google TPU



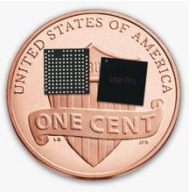


TPU v2 - 4 chips, 2 cores per chip



TPU v3 - 4 chips, 2 cores per chip

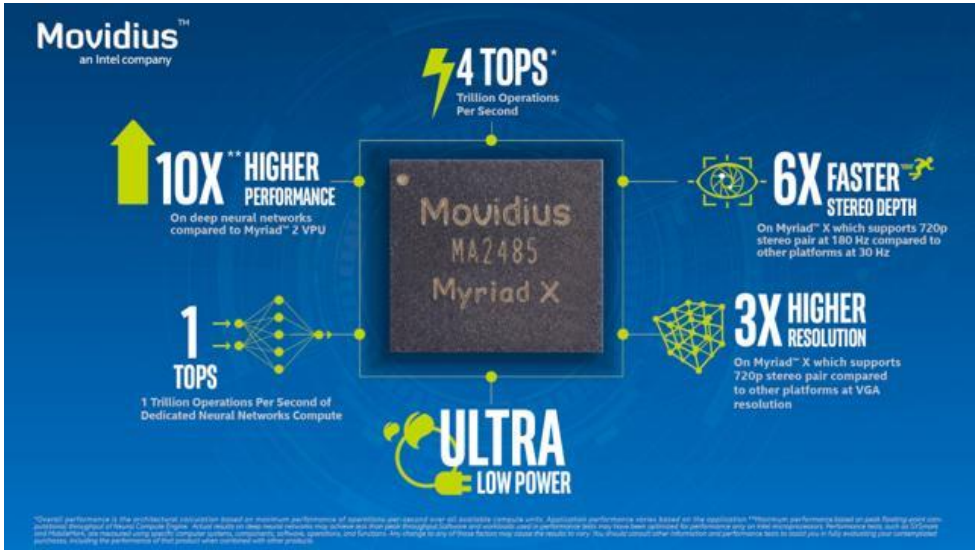
- 歷經四代演進，為資料中心專用的神經網路加速晶片。
- 搭配TensorFlow框架開發，具有更佳效能。
- Google Colab有提供免費算力，方便測試。

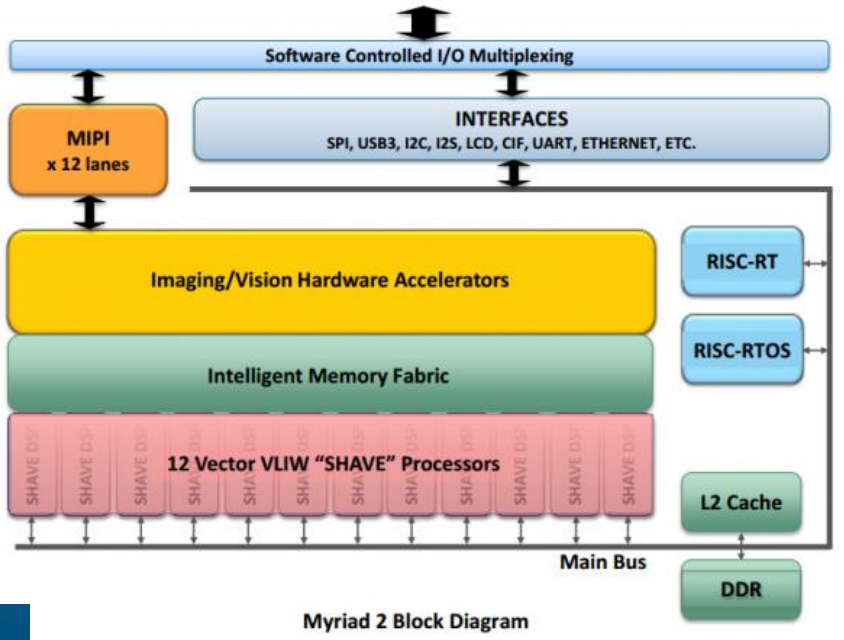


Edge TPU
(Int8)

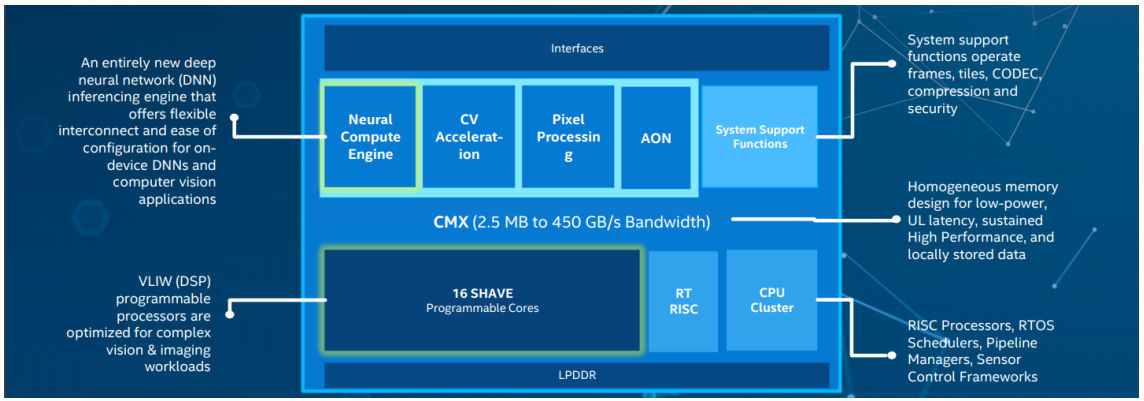
2.2.4 NPU–Intel (Movidius)

Myriad X (MA2485), 又稱VPU。





Myriad 2 Block Diagram



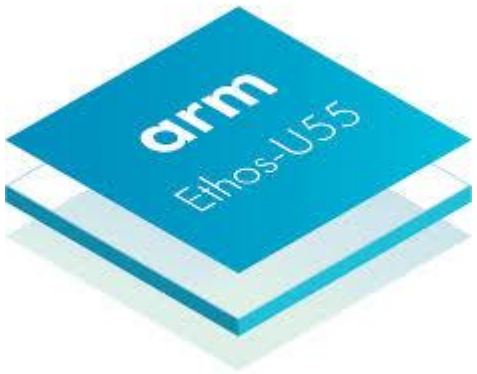
4 TOPS
16 x 128bit VLIW Vector Processors
16 x MIPI Lanes
4GB LPDDR4

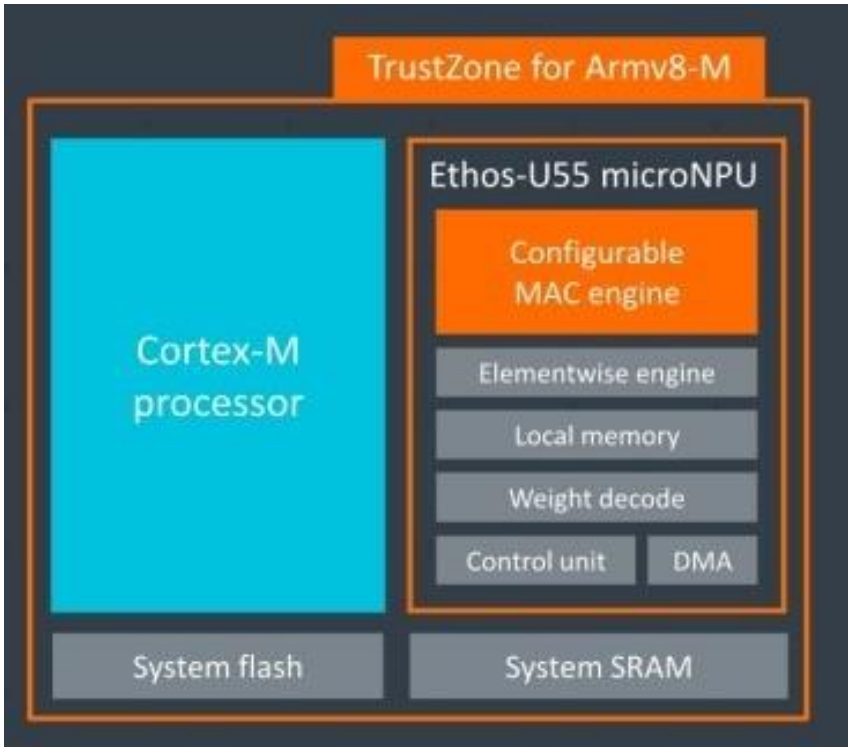
2.2.4 NPU—耐能Kneron

- KL520
  - Cortex-M4 x 2
  - 32 MB Flash
  - 64 MB LDDR2
  - 0.35TOPS / 0.5W
  - 支援ONNX, TensorFlow, Keras, Caffe

2.2.4 NPU–ARM

- ARM Ethos U55
  - 可搭配多款MCU(Cortex-M55, M33, M7, M4)使用
  - 32, 64, 128 or 256 unit MAC引擎
  - 具DMA功能可快速存取模型權重值





2.2.4 NPU—嘉楠

- 勘智K210
  - RISC-V 64bit CPU x 2
  - 1 TOPS (KPU)
  - 300 mW
  - 支援 TensorFlow / Keras / DarkNet(TinyYOLOv2)

嘉楠科技
Canaan Inc.

資料來源：https://canaan-creative.com/product/kendryteai

2.2.4 NPU—瑞芯微

- RK1808
  - Dual Cortex-A35 1.6GHz
  - 3 / 0.3 / 0.1 TOPs (INT8, INT16, FP16)
  - Support OpenCL/VX, TensorFlow, Caffe, ONNX, Darknet

Rockchip
瑞芯微电子

2.2.5 ASIC—供應商

- **華為海思** 昇騰(Ascend) 310 / 910
- **紫光展銳** 虎賁T710
- **地平線** 征程2/3, 旭日2/3
- **寒武紀** 思元 290/270/100/200
- **比特大陸** 算丰BM1680 / 1682 / 1684 / 1880
- **遂原科技** 燧思, 雲燧T10 / T11 / i10
- **雲天勵飛** DeepEye1000
- **全志科技** R329, R818
- **瑞芯微** RK3399Pro / RK1808
- **鯤雲科技** CAISA
- **依圖科技** 求索
- **啟英泰倫** CI100X / 110X / 112X
- **知存科技** MemCore001 / 001P / 101 / 201
- **黑芝麻智能** A1000
- **深聰智能** 太行TH1520
- **肇觀電子** N / D / V系列
- **天數智芯** BI
- **探境科技** 音旋風611 / 612 / 621 / 631 / 711 / 712, Imagist851
- **嘉楠科技** 堪智K210 / 510
- **雲知聲** 蜂鳥
- **清微智能** TX101 / 210 / 510 (可重構芯片)
- **酷芯微** AR9000 / 9200
- **杭州國芯** GX8002 / 8010 / 8009 / 8008 / 8001
- **耐能科技** KL520 / 720
- **平頭哥** 含光800
- **百度** 崑崙1 / 2
- **北京君正** T02 / T31

資料來源：https://www.esmchina.com/news/7491.html

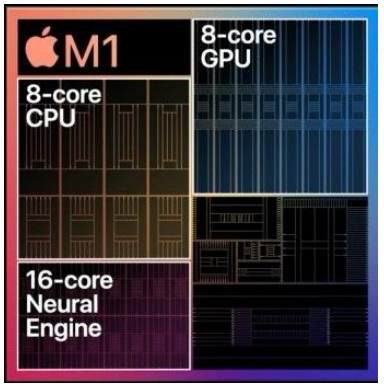
2.2.5 ASIC—應用情境

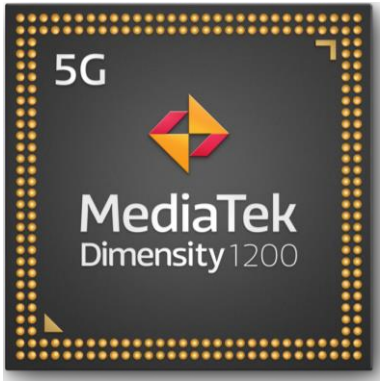
- 智慧音箱（語音命令）
- 自駕車
- 安全監控
- 人臉 / 語音辨識
- 智慧製造 / 倉儲 / 教育
- 智慧城市 / 家庭 / 生活
- 智慧醫療 / 交通 / 金融
- **雲端機房訓練 / 推理**
- 視頻分析
- 物聯網(AIoT)
- 機器人視覺
- 可穿戴裝置

資料來源：https://www.esmchina.com/news/7491.html

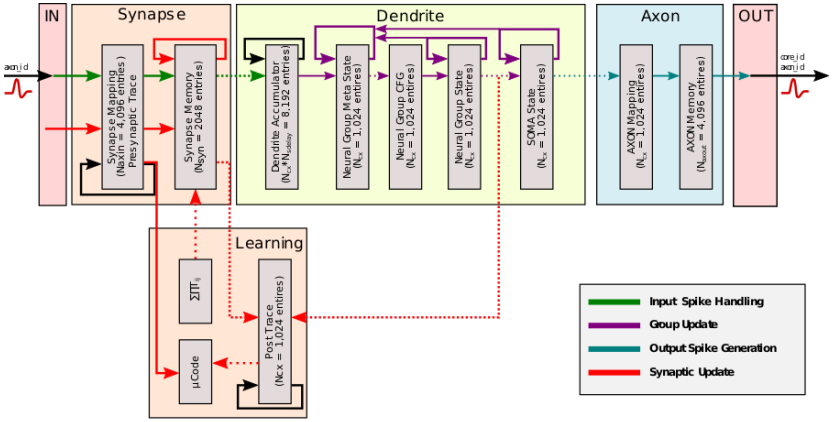
2.2.6 其它類型—系統級晶片SoC

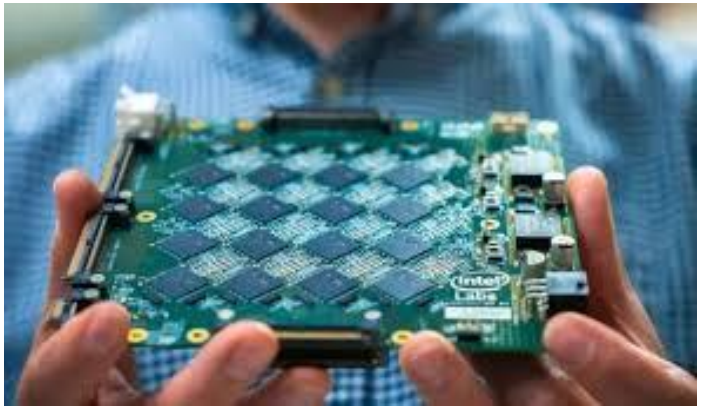
- 智慧手機/筆電晶片
  - 通常以ARM系列CPU為主要核心。另外會搭配GPU，NPU，DSP，ISP等加速計算功能。
  - 可搭配 Android NN 框架進行AI模型推論。
- 常見具AI處理器SoC晶片
  - 高通(驍龍 888)
  - 蘋果(Apple M1)
  - 聯發科(天璣 1200)
  - 華為海思(麒麟 990)





2.2.6 其它類型—類腦晶片

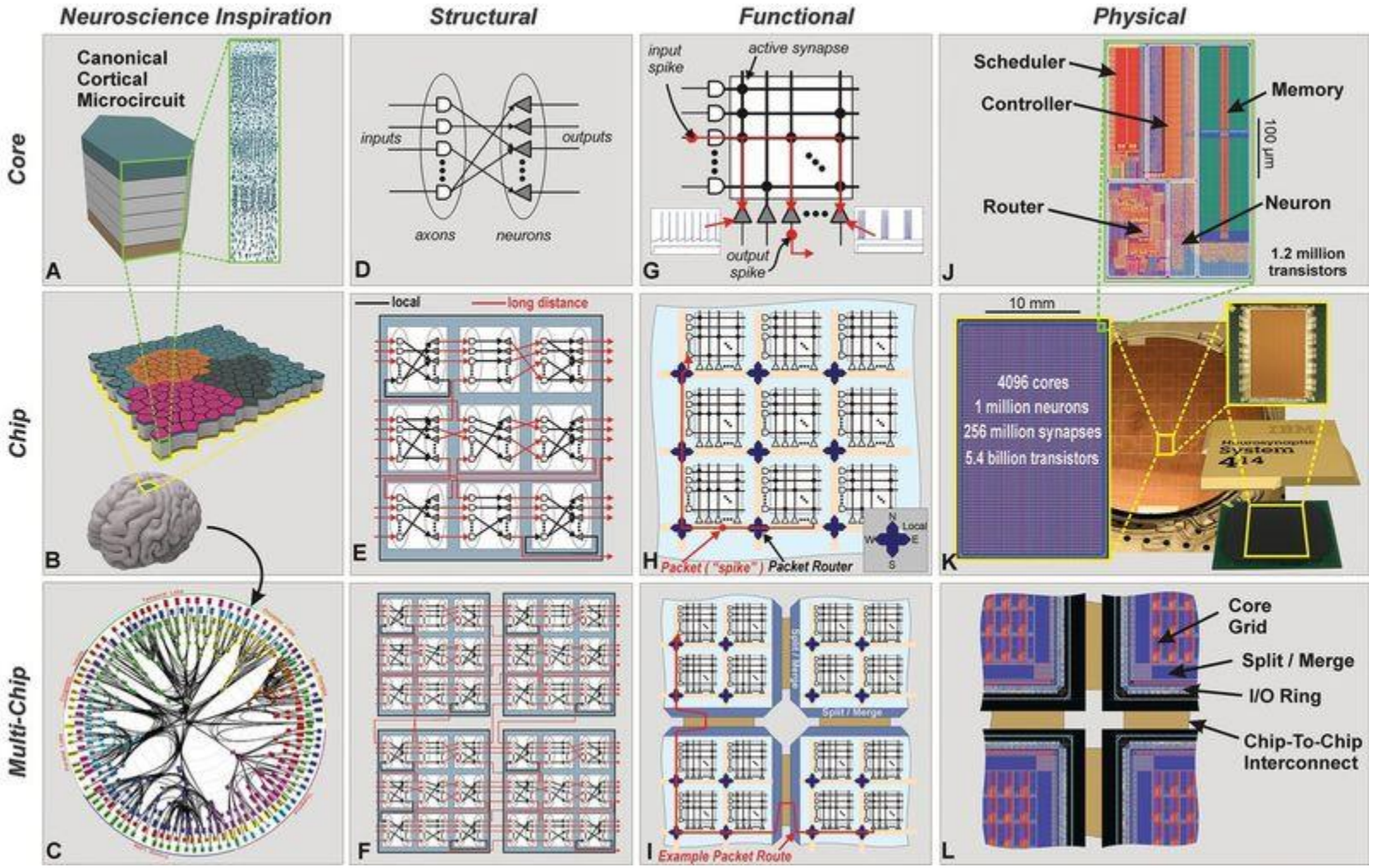




INTEL Pohoiki Beach
(64 x Loihi Chip)

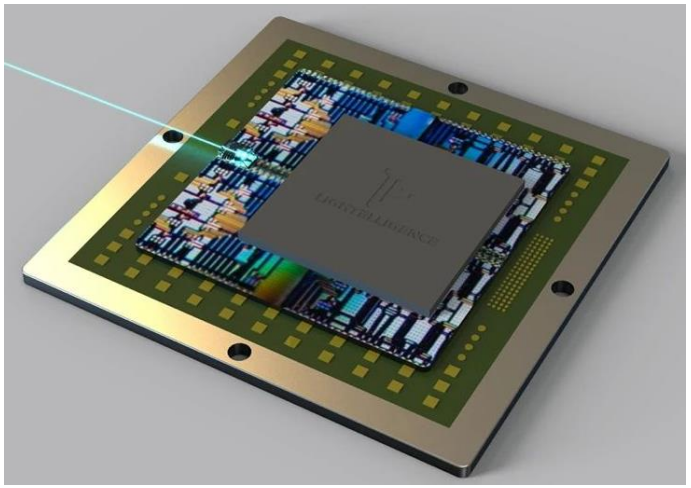
- **Neuromorphic 神經形態**
  - 數位式
    - IBM TrueNorth
    - INTEL Loihi
    - SpiNNaker
  - 類比式
    - Neurogrid
    - BrainScales
    - ROLLS
  - 新材料
    - Memristor

2.2.6 其它類型–IBM TrueNorth



資料來源：https://www.greencarcongress.com/2014/08/20140808-truenorth.html

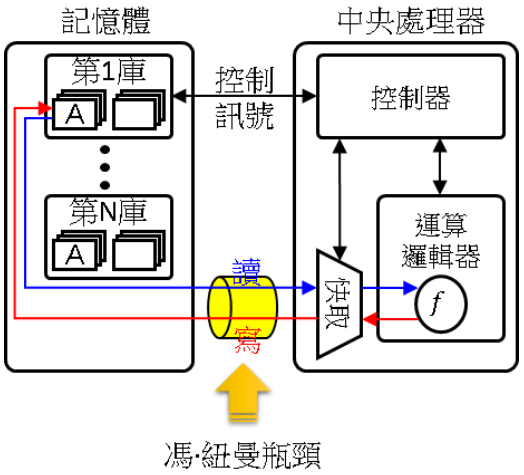
2.2.6 其它類型—光子晶片

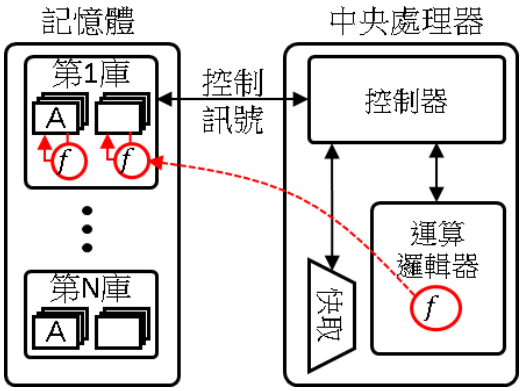


- Lightmatter
  - MIT獨立出新創公司
  - 利用矽光子和微機電製程技術，以毫瓦等級雷射光供電，能以接近光速執行矩陣向量乘法。
- 曦智lightelligence
  - 大陸團隊，百度投資
  - 和Lightmatter皆為MIT技術，技術類似。

資料來源：https://www.eettaiwan.com/20201006nt31-optical-compute-promises-game-changing-ai-performance/
資料來源：https://www.lightelligence.co/

2.2.6 其它類型—記憶體內記計算

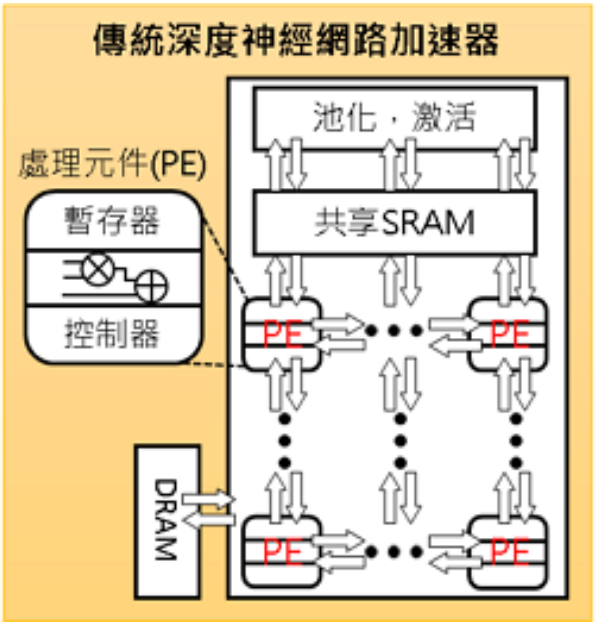


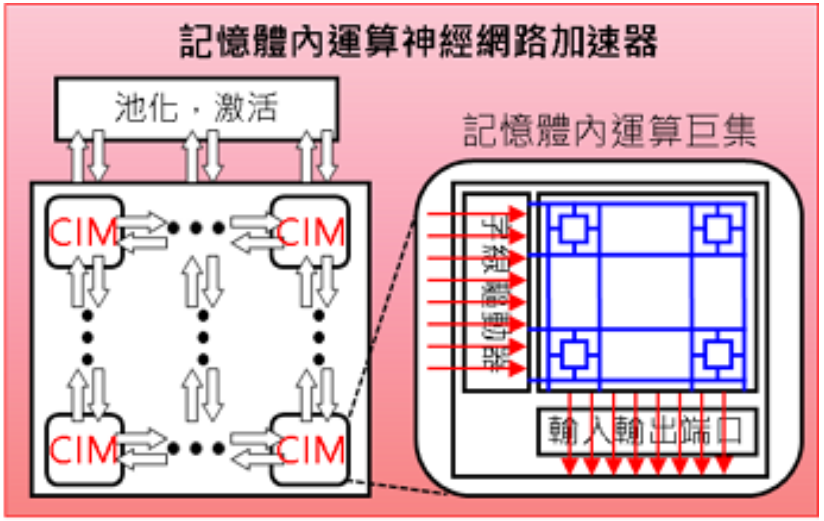


In-Memory Computing 存算一體、存內計算

減少記憶體到計算單元距離來加速運算

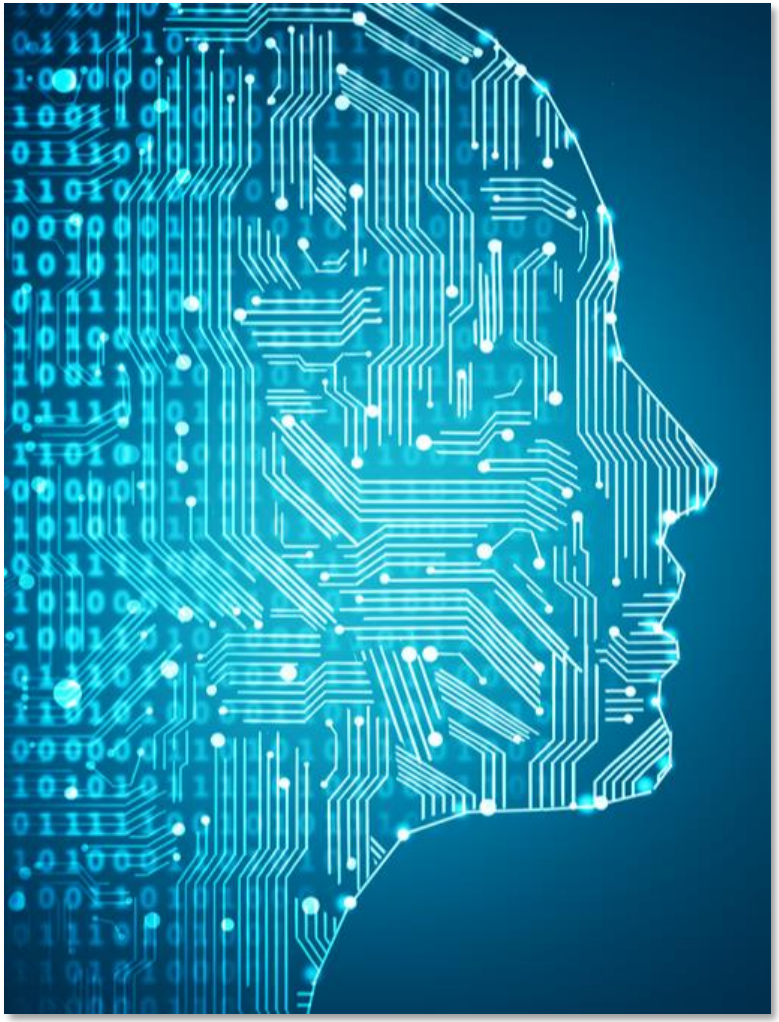
* SRAM (靜態隨機存取記憶體 / 電晶體式)
* MRAM (磁阻式) * RRAM (可變電阻式)





資料來源：
https://ictjournal.itri.org.tw/content/Messagess/contents.aspx?PView=1&SiteID=654246032665636316&MmmID=6543044320616444 11&SSize=10&MSID=1037365734470512074

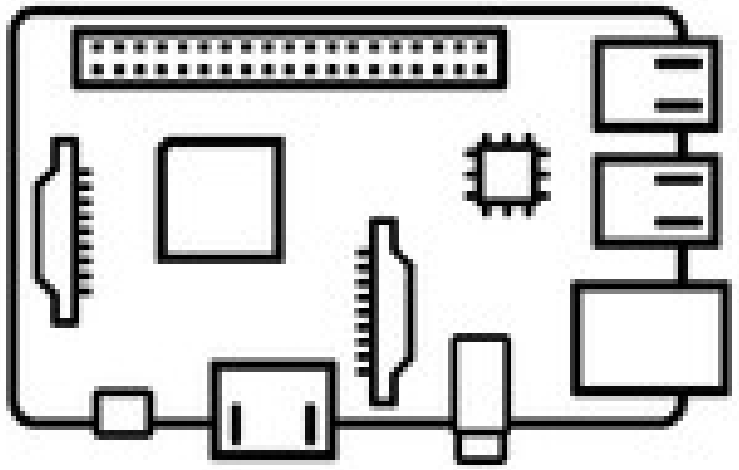
2.3 開發板類型



- 單板微電腦
- USB加速棒
- 模組板卡
- 整合型裝置

2.3.1 單板微電腦

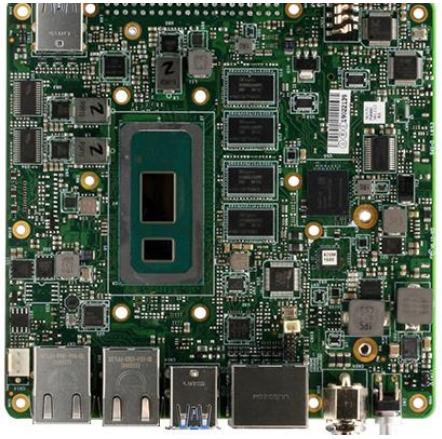
- Intel
- Nvidia
- Google



- ARM
  - Raspberry Pi
  - Asus Tinker
  - Rockchip
  - Arduino
  - Banan Pi
- FPGA

2.3.1 單板微電腦–INTEL

AAEON UP Squared
(Pentium N4200 / Celeron N3350)



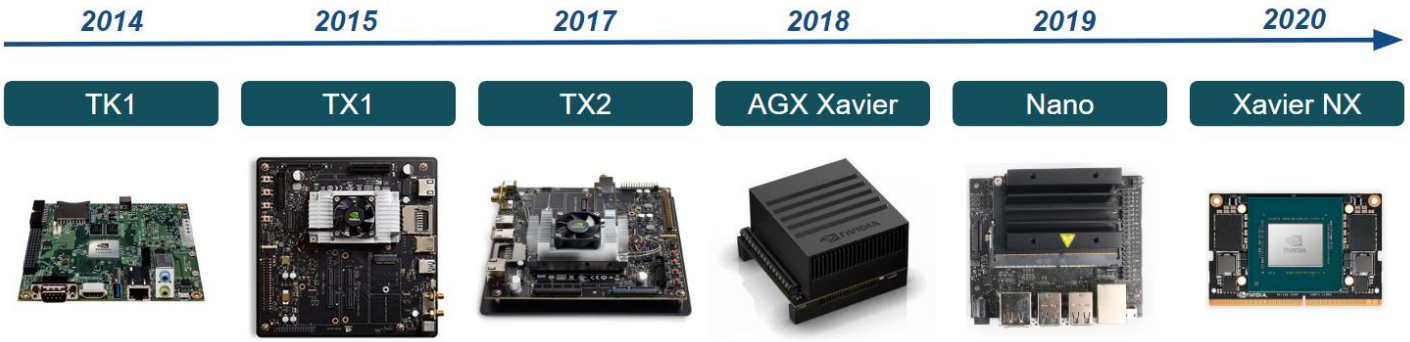
AAEON UP Extrme
(Core i7/i5/i3/Celeron)

AAEON COM-TGUC6
(Tiget Lake UP3 with GPU Iris Xe)

Intel Arria 10 SoC Board
(FPGA Arria 10)

AAEON BOXER-8320AI
(Core i3-6100U + Dual Myriad VPUs)

2.3.1 單板微電腦—NVIDIA



2.3.1 單板微電腦—GOOGLE

Google Coral (Edge TPU) Series

Asus Tinker Edge T

Dev Board

Dev Board Mini

2.3.1 單板微電腦—樹莓派

ARM Based CPU
+Broadcom GPU
VideoCore IV

Pi 0	ARMv6z(32bit)
Pi 3B	ARMv8-A(32/64bit)
Pi 3A+	ARMv8-A(32/64bit)
Pi 4	ARMv8-A(32/64bit) (VFPv4 + NEON)

Pi 3A+ (4 * Cortex-A53 1.4GHz)

Pi 0 (Arm11 1GHz)

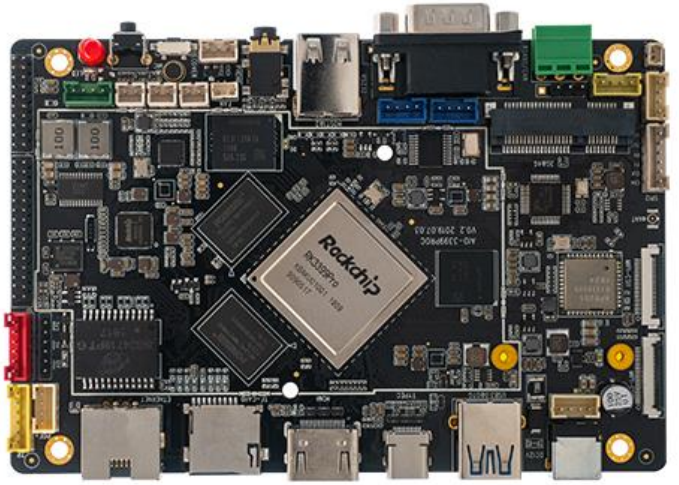
Pi 3B (4 * Cortex-A53 1.4GHz)

Pi 4B (4 * Cortex-A72 1.5GHz)

2.3.1 單板微電腦–ARM開發板1

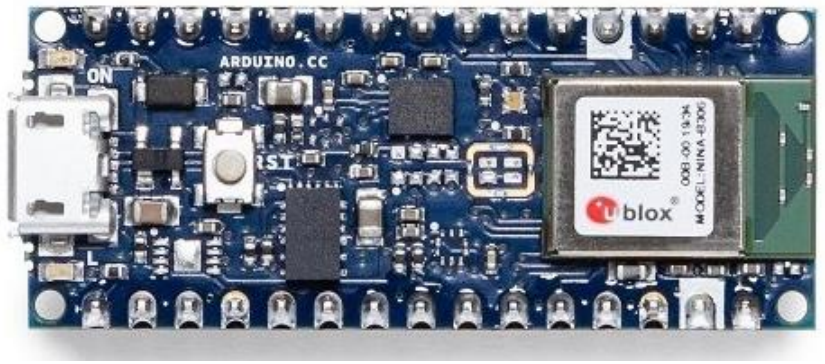
- **ASUS Tinker Edge T**
  - NXP i.MX 8M(Cortex-A53 * 4 + Cortex-M4 * 1) + Google Edge TPU
  - 1GB RAM / 8GB eMMC
  - 4 TOPS(INT8)

- **Rockchip RK3399Pro**
  - Cortex-A72 * 2 + Cortex-A53 * 4
  - NPU 3 TOPS(INT8/16)
  - GPU Mali-T860MP4

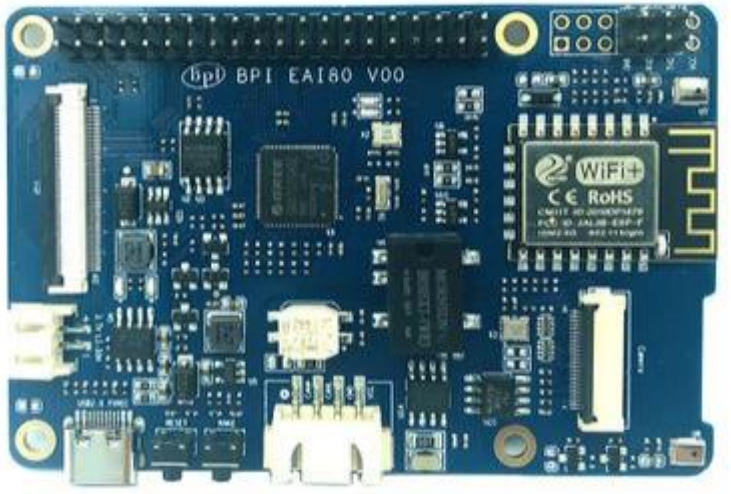


2.3.1 單板微電腦–ARM開發板2

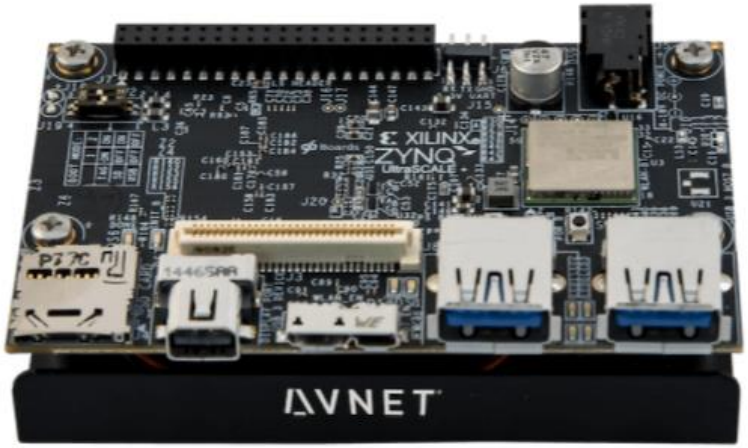
- **Arduino**
  - Nano 33 BLE
  - nRF52840 (Cortex-M4)
  - 64 MHz
  - 1MB Flash / 256KB SRAM



- **Banana Pi**
  - BPI-EAI80 AIoT
  - Dual-Cortex M4F@200MHz
  - NPU 300GOPS
  - SDRAM 8MB



2.3.1 單板微電腦–FPGA



- AVNET FPGA開發板
  - Ultra96-V2
  - Xilinx Zynq UltraScale+ MPSoC ZU3EG A484
  - 2GB RAM / 16 GB SD

- Microsemi FPGA開發板
  - M2S-HELLO-FPGA-KIT
  - SmartFusion 2 FPGA + Arm Cortex –M3

2.3.2 USB加速棒

- **Intel NCS 2**
  - Movidius VPU MA2485
  - USB 3.0 / 2.0
  - 4 TOPS
  - Support OpenVINO

- **Google Coral USB**
  - Edge TPU
  - NXP i.MX 8M SoC (quad Cortex-A53, Cortex-M4F)
  - 4 TOPS(INT8), 2 TOPS/W

2.3.3 模組板卡—SOM

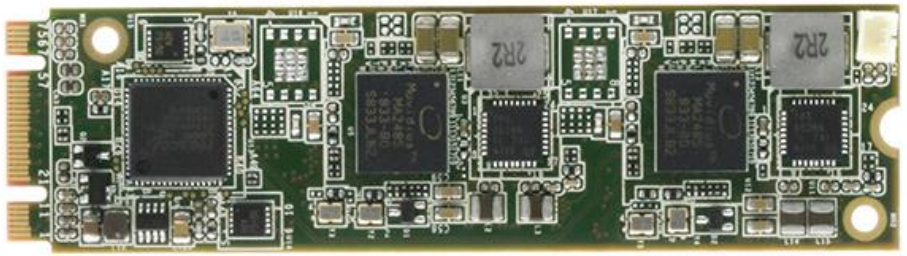
- **RockChip SOM**
  - RK3399Pro
  - Cortex-A72 * 2 + Cortex-A53 * 4
  - NPU 3 TOPS(INT8/16)
  - GPU Mali-T860MP4
- **Google Coral SOM**
  - Edge TPU
  - NXP i.MX 8M SoC (quad Cortex-A53, Cortex-M4F)
  - 4 TOPS(INT8), 2 TOPS/W
  - 1GB RAM / 8GB eMMC

2.3.3 模組板卡—M.2 / mPCIE

AI CORE XM 2280

- Intel Movidius Myriad X VPU, MA2485 * 2
- Support Tensorflow, Caffe, MXNET
- M.2 2280 B+M Key
- Support Intel OpenVINO toolkit

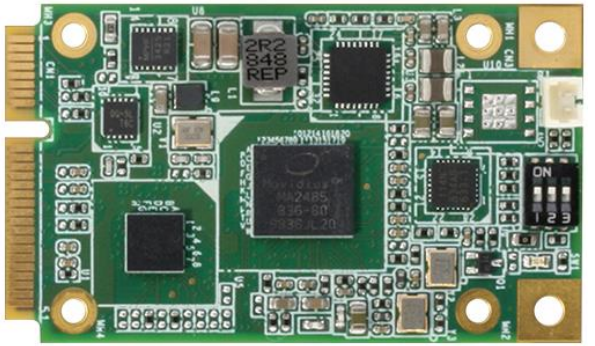
AAEON



AI CORE X

- Intel Movidius Myriad X VPU, MA2485 * 1
- Support Tensorflow, Caffe, MXNET
- mini PCIE (mPCIe)
- Support Intel OpenVINO toolkit

AAEON



2.3.4 整合型裝置—智慧音箱

資料來源：https://img.itw01.com/images/2018/03/09/19/5819_pan6ca_3B4BEHN.jpeg!r800x0.jpg

2.3.4 整合型裝置—智慧攝影機

ADLINK NEON-1000-MDX
Intel Atom E3930 + Movidius VPU

ADLINK NEON-2000-JNX
NVIDIA Jetson Xavier NX-based

OpenMV Cam H7
ARM Cortex M7

Himax WE-I Plus EVB
ARC 32-bit EM9D DSP with FPU

2.4 硬體選用評估

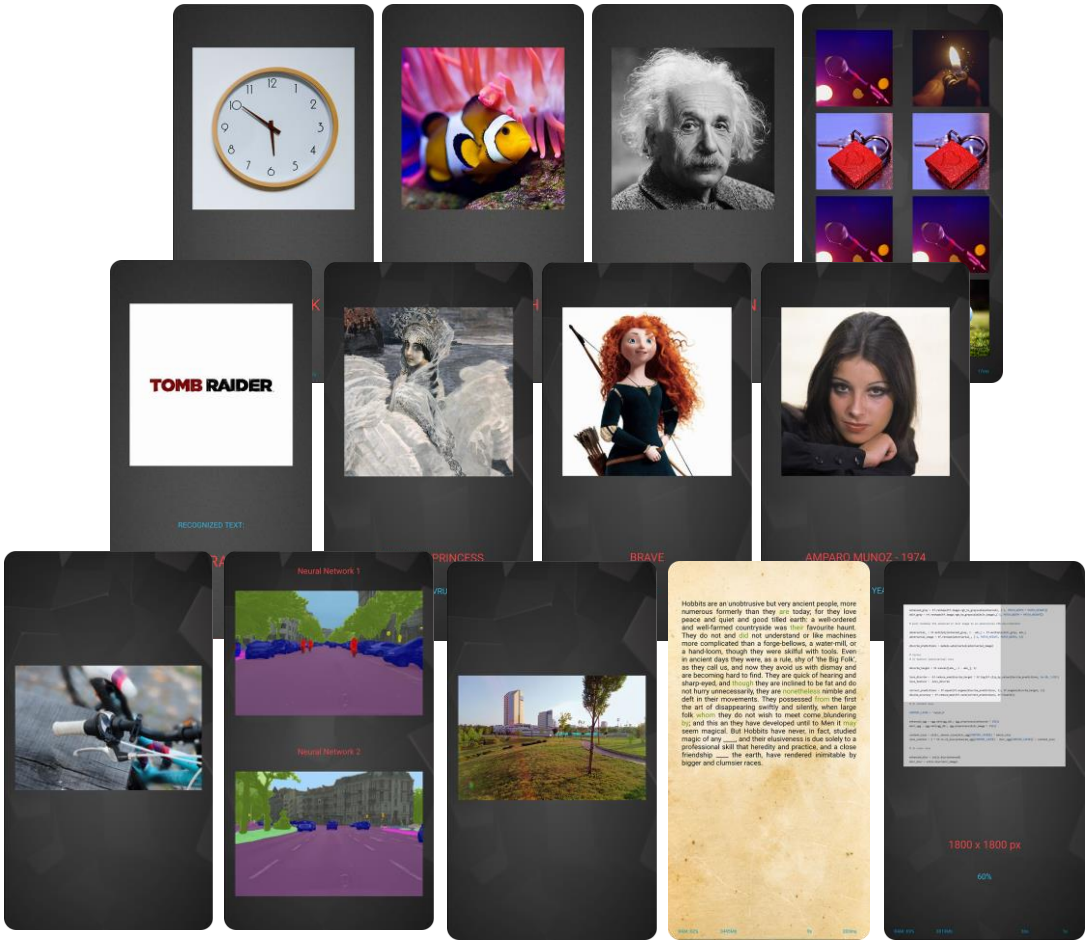
- 運算效能
- 開發工具
- 應用情境
- 週邊擴展

2.4.1 運算效能—主要指標

- 主要指標
  - TOPS（每秒運算次數, 單位Tera,10^{12}）
  - TOPS / W （每瓦運算次數）
  - 算力成本（每塊錢得到之TOPS）
  - 數值精度（INT8 / INT16 / FP16 / FP32）
  - Top 1 (Top 5)運算精確度（Accuracy）
  - 均值平均精確度(mean Average Precision, mAP)
  - 推論速度（Frame per Second, FPS）
  - 批量/延遲（Batch Size / Delay）
  - 整體功耗（Power Consumption, W）

2.4.1 運算效能—AI-Benchmark

- ETH 手機晶片AI測試
  - Object Classification (MobileNet-V2 / Inception-V3 / MobileNet-V2)
  - Face Recognition (MobileNet-V3 Large-M)
  - Optical Character Recognition(CRNN / Bi-LSTM)
  - Image Deblurring (PyNET-Mini)
  - Image Super-Resolution(VGG-19 / SRGAN)
  - Bokeh Simulation (U-Net)
  - Semantic Segmentation (DeepLab-V3+)
  - Photo Enhancement (DPED-ResNet)
  - Text Completion (Static RNN / LSTM)
  - Memory Limits (SRCNN 9-5-5)



資料來源：https://ai-benchmark.com/

2.4.1 運算效能—ML COMMONS

Training

MLPerf

已併入

ML Commons

Area	Benchmark	Dataset	Quality Target	Reference Implementation Model
Vision	Image classification	ImageNet	75.90% classification	ResNet-50 v1.5
Vision	Object detection (light weight)	COCO	23.0% mAP	SSD
Vision	Object detection (heavy weight)	COCO	0.377 Box min AP and 0.339 Mask min AP	Mask R-CNN
Language	Translation (recurrent)	WMT English-German	24.0 Sacre BLEU	NMT
Language	Translation (non-recurrent)	WMT English-German	25.00 BLEU	Transformer
Language	NLP	Wikipedia 2020/01/01	0.712 Mask-LM accuracy	BERT
Commerce	Recommendation	1TB Click Logs	0.8025 AUC	DLRM
Research	Reinforcement learning	Go	50% win rate vs. checkpoint	Mini Go (based on Alpha Go paper)

Inference: Edge

Area	Task	Model	Dataset	QSL Size	Quality	Multi-stream latency constraint
Vision	Image classification	Resnet50-v1.5	ImageNet (224x224)	1024	99% of FP32 (76.46%)	50 ms
Vision	Object detection (large)	SSD-ResNet34	COCO (1200x1200)	64	99% of FP32 (0.20 mAP)	66 ms
Vision	Object detection (small)	SSD-MobileNets-v1	COCO (300x300)	256	99% of FP32 (0.22 mAP)	50 ms
Vision	Medical image segmentation	3D UNET	BraTS 2019 (224x224x160)	16	99% of FP32 and 99.9% of FP32 (0.85300 mean DICE score)	N/A
Speech	Speech-to-text	RNNT	Librispeech dev-clean (samples < 15 seconds)	2513	99% of FP32 (1 - WER, where WER=7.452253714852645%)	N/A
Language	Language processing	BERT	SQuAD v1.1 (max_seq_len=384)	10833	99% of FP32 (f1_score=90.874%)	N/A

資料來源：https://mlcommons.org/

2.4.2 開發工具–語言與工具

- 主要程式語言
  - C / C++
  - Python
  - ASM
- 作業系統
  - Linux (Raspbain, Ubuntu)
  - Windows / Mac
  - Android / iOS
  - ROS, Mbed, RTOS….
- 平行處理語言
  - GPGPU
  - CUDA
  - OpenCL
  - OpenVX
- 開源工具
  - OpenCV
  - Google Colab
  - Intel OpenVINO

2.4.2 開發工具—框架


TensorFlow
2.0
Keras
Chainer
Marvin
neon
tiny-dnn
Caffe
USI/SUPSI
IDSIA
Microsoft
CNTK
ML
CoreML
Torch
Paddle
DL4J
ONNX
theano
BigDL
APACHE
Spark
Caffe2
dy/net
PYTORCH
mxnet
TensorFlow Lite
TensorRT
OpenVINO
arm
CMSIS-NN

2.4.3 應用情境


單晶片
單板微電腦
桌機/伺服器
MCU
SBC
PC
運算速度 /
平均功耗 /
單位成本
感測分析
語音命令
運動偵測
影像分類
物件偵測
語音辨識
模型訓練
影像分割
自然語言

2.4.4 週邊擴展


攝像頭(光源、鏡頭)
專用型、USB Webcam、立體攝像頭、深度感測器(RGB-D)
通用型輸出入控制點(GPIO)
輸入：觸發信號、開關(按鍵)、感測器...
輸出：結果信號、光源控制、馬達動作...
通訊：SPI, I2C, UART ...
加速硬件
Intel Movidius USB神經運算棒
計算用主機
單/多核CPU, GPU, SDRAM, SD卡(Flash)
通訊介面
WiFi, BT, USB, Ethernet
作業系統
驅動程式
OpenCV
應用程式
(樹莓派3為例)
影音介面
HDMI, VGA, LCD, OLED, 麥克風、耳機

小結

- **基本運算原理**小節中介紹了基本運算原理，包括卷積神經網路、 數字表示系統、矩陣 / 張量運算及平行 / 並行運算 。
- **加速運算晶片**小節中介紹各種晶片型式，包括CPU, GPU, VPU, FPGA, ASIC及其它類型晶片。
- **開發板類型**小節中介紹各種單板微電腦、USB加速棒、不同介面模組板卡及智慧音箱和智慧攝影機。
- **硬體選用評估**小節中分別介紹和運算效能相關的名詞、各種開發工具框架，並延至應用情境及未來如何將週邊擴展。

參考文獻

- 許哲豪，“人工智慧AI晶片與Maker創意接軌”

http://omnixri.blogspot.com/2018/05/aimaker.html

- 許哲豪，“AI 邊緣運算即將掀起風潮”

http://omnixri.blogspot.com/2018/07/aicolumnai.html

- 許哲豪，“樹莓派GPU的算力釋放”

http://omnixri.blogspot.com/2018/07/gpu.html

- 許哲豪，“AI晶片如何評比效能”

http://omnixri.blogspot.com/2019/03/ai.html